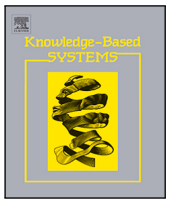
Knowledge-Based Systems 310 (2025) 112926

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys



Disentangle Representation Learning with Excluding Confounding Bias for causal effect estimation

Dianlong You [a,b], Dongyan Wang [a,b], Bingxin Liu [a,b], Xiaoyi Ge [a,b], Di Wu [c,*], Xindong Wu [d]

[a] School of Information Science and Engineering, Yanshan University, Qinhuangdao, 066004, Hebei, China
[b] The Key Laboratory for software engineering of Hebei Province, Yanshan University, Qinhuangdao, Hebei, 066004, China
[c] The College of Computer and Information Science, Southwest University, Chongqing, 400715, China
[d] The Key Laboratory of Knowledge Engineering with Big Data (Hefei University of Technology), Ministry of Education, Hefei, 230009, China

ARTICLE INFO

Keywords:
Disentangle representation learning
Confounding bias
Causal effect estimation
Treatment and effect

ABSTRACT

Causal effect estimation aims to discover the impact of treatments on effects from observational data. Existing approaches suffer from effectively overcoming confounding bias, in which the core challenge is to disentangle confounders from treatment and effect variables simultaneously. To this end, we propose a novel Disentangle Representation Learning approach with Excluding Confounding Bias mechanism (DRL_{ECB}) for causal effect estimation with the three-fold ideas: (1) Disentangling treatment variables into instrumental, confounders and risk factors by maximizing and minimizing mutual information (MI); (2) Balancing confounding representations to eliminate confounding bias by reweighting on treatment and control groups; (3) Constructing outcome regression networks with confounding and risk representations to predict outcomes and summarizing the total loss to estimate causal effects. We conducted extensive experiments on benchmark and synthetic datasets. The results demonstrate that DRL_{ECB} performs significantly better than its rivals. The code is open-source and publicly available at https://github.com/youdianlong/DRLECB.

1. Introduction

Causal effect estimation plays an essential role in decision-making process, and interpretable statistical analysis [1], remaining wide applications across diverse domains including health care [2–5], policy evaluation [6–8], machine intelligence [9–12] and online advertising [13–16]. For example, in policy evaluation, whether or not the vocational training plan formulated by decision-makers upgrades employment prospects [17]; how do companies adjust coupon allocation plans to increase purchase rates as consumers grow [18]? After a treatment cycle, whether or not new medicines the patient is taking improve diabetes symptoms. Classic approach, e.g., Randomized Controlled Trial (RCT) [19,20], commonly assigns treatments (i.e., medical procedures) to random units (i.e., patients) [21] to discover causal effects. Admittedly, this is usually long-term, expensive [22], and even unethical [23]. Therefore, mining causal effects from existing observable data becomes a popular and challenging paradigm [24,25]. Coming with it, confounding bias is a dilemma that needs to be addressed during learning due to the effects of treatment variables. Because the change could lead to confounding bias ($P(T|X) \neq P(T)$) [26]. Fig. 1 shows a relationship between COVID-19 and mortality. Generally, COVID-19 can result in increasing mortality. Meanwhile, the elderly have a higher mortality rate and are more susceptible to infection. The confounders 'age' affects both mortality and infection rates simultaneously. Therefore, we need to dominate confounders to eliminate bias.

Some methods partially solve the above confounding bias problem, including Counterfactual Regression (CFR-MMD and CFR-WASS) [27], CounterFactual Regression with Importance Sampling Weights (CFR-ISW) [28], Similarity preserved Individual Treatment Effect (SITE) [29], Double Robust Representation Learning (DRRL) [30], Disentangled Representations for CounterFactual Regression (DR-CFR) [31], Treatment Effect Disentangling Variational AutoEncoder (TEDVAE) [32], Disentangled Representations for CounterFactual Regression (DeR-CFR) [21], Disentangled Representations for Counterfactual Regression via Mutual Information Minimization (MIM-DRCFR) [25], Automatic Instrumental Variable decomposition (AutoIV) [33]. Among them, (1) Representation-based learning approaches, e.g., CFR-MMD, CFR-WASS [27], CFR-ISW [28], SITE [29] and DRRL [30], consider treatment variables X as confounders, by balancing the confounders and thereby reducing the confounding bias. But, do not distinguish confounders and non-confounders in treatment variables X. Balancing non-confounders will reduce the accuracy and credibility of the prediction results. (2)

* Corresponding author.
E-mail address: wudi1986@swu.edu.cn (D. Wu).

https://doi.org/10.1016/j.knosys.2024.112926
Received 13 September 2024; Received in revised form 5 December 2024; Accepted 23 December 2024
Available online 30 December 2024

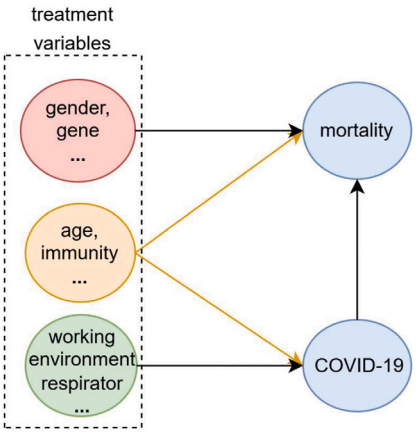


Fig. 1. In treatment variables, gender and gene only affect mortality; age and immunity affect both mortality and COVID-19; working environment and respirator only affect COVID-19.

Decomposition-based approaches, e.g., DR-CFR [31], TEDVAE [32], DeR-CFR [21], MIM-DRCFR [25] and AutoIV [33] disentangle confounders from treatment variables X. However, DR-CFR [31] and AutoIV [33] fails to obtain independent confounders; TEDVAE [32] lacks special mechanisms for excluding confounding bias; DeR-CFR [21] is not suitable for high dimensional and large-scale data due to its complex structure; The information learned by the multi-task representation in MIM-DRCFR [25] may be too abstract or generic, which may lead to ineffective disentangling. Motivated by this situation, this paper explores a new Disentangle Representation Learning model with Excluding Confounding Bias mechanism to discompose treatment variables for causal effect estimation, termed DRL_{ECB}, based on the assumptions of Stable Unit Treatment Value (SUTV), Ignorability, and Positivity (Section 3 for details). The challenges of implementing DRL_{ECB} lie in three aspects: (1) How to disentangle the treatment variables for mining confounders? (2) How to exclude confounding bias from treatment and control groups? (3) How to effectively estimate causal effects? Our DRL_{ECB} model argues a three-fold idea: (1) Inspired by DR-CFR [31], it disentangles the treatment variables X into three independent parts, i.e., $\Gamma(X), \Delta(X), \Upsilon(X)$, by maximizing and minimizing mutual information (MI). As shown in Fig. 2, (i) disentangling $\Gamma(X)$ from X to make that $\Gamma(X)$ is correlated with T and conditionally independent of Y given treatment T; (ii) disentangling $\Delta(X)$ from X, by making $\Delta(X)$ is correlated with T and Y; (iii) disentangling $\Upsilon(X)$ from X, to make $\Upsilon(X)$ correlated with Y and uncorrelated with T; (2) It achieves a balance of confounding representation by reweighting treatment and control groups; (3) It designs network regression models to generate predict outcomes and construct a total loss to achieve treatment effect estimations with confounding and risk representations, respectively.

The main contributions in this paper are as follows:

- Our proposed DRL_{ECB} model achieves decomposition via MI rather than MMD to obtain the confounding representation based on the correlation between the three factors $\Gamma(X), \Delta(X), \Upsilon(X)$ and T, Y, which provides a foundation for eliminating the confounding bias due to the confounders.
- To eliminate the confounding bias, we balance the confounding representation between treatment and control groups by reweighting, which reduces the distribution distance of the confounding representation between treatment and control groups.
- To estimate causal effects, we first construct two regression networks with risk and balanced confounding representations to predict outcomes in treatment and control groups, respectively. Then, construct a total loss composed of variational approximation, MI constraints, balancing, and prediction losses to estimate causal effects, which can enormously disentangle confounders and exclude bias.

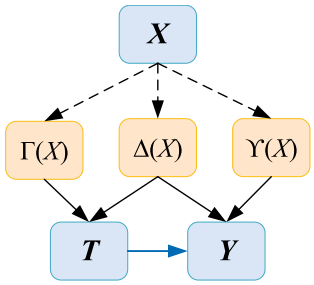


Fig. 2. Inspired by DR-CFR [31], treatment variables X are disentangled instrumental factors $\Gamma(X)$, confounders $\Delta(X)$ and risk factors $\Upsilon(X)$. $\Gamma(X)$ and $\Delta(X)$ affect treatment T; $\Delta(X)$ and $\Upsilon(X)$ affect outcome Y.

The remainder of the paper is organized as follows. Section 2 reviews some existing work related to our model; Section 3 describes DRL_{ECB} preliminaries; Section 4 describes our DRL_{ECB} model in detail; Section 5 describes the theoretical guarantees of the DRL_{ECB} model; Section 6 presents our main experimental results concerning the theoretical guarantees; Section 7 concludes the study.

2. Related work

2.1. Representation-based approach

Representation learning-based approaches focus on adjusting confounding variables by learning a balanced representation of all covariates [34]. In the CFR-MMD and CFR-WASS [27] models, Maximum Mean Discrepancy (MMD) or Wasserstein (WASS) distance are used to reduce the distances between the distributions of different treatment groups in the representation space, respectively. When the learned representation dimension is high, using that representation as input may lose the impact of different treatments on the effect. CFR-ISW [28] presents a context-aware importance sampling weighting scheme based on representation learning to reduce bias. The above methods of effect estimation balance the distribution of treatment and control groups from a global perspective, ignoring the local similarity information between them. So, SITE [29] preserves local similarity information in the representation space through the midpoint distance minimization strategy (MPDM) and the position-dependent depth criterion (PDDM), meanwhile balancing the data distributions. DRRL [30] is the first to combine the Doubly Robust (DR) estimator with representation learning, learning weights through the entropy balance method to minimize the Johnson-Shannon (JS) divergence of the representation between treatment and control groups, and using DR to make robust and effective estimates of the treatment effect [30]. However, these approaches learn representation from all treatment variables to achieve balanced representation between treatment and control groups but neglect to distinguish between non-confounders, resulting in additional errors.

2.2. Decomposition-based approach

Decomposition-based algorithms mainly disentangle the treatment variables X into confounders and non-confounders in estimating treatment effects. (1) DR-CFR [31] minimizes distribution discrepancy (disc) using MMD to disentangle treatment variables into roughly three parts, in which decomposed instrumental factors and confounders cannot be effectively separated. (2) DeR-CFR [21], similarly to the DR-CFR [31]

model, still uses the minimization disc method to disentangle the treatment variables X. To achieve a more precise decomposition, DeR-CFR [21] maximizes the decomposition among variables by a regression model and deep orthogonal regularize, leading to high time costs. (3) TEDVAE [32] utilizes a variational autocoder (VAE) to generative disentangled model for learning instrumental, confounding, and risk factors. Although the model disentanglement yields confounders, it cannot effectively reduce the confounding bias owing to lacking manipulation of confounders. Additionally, the generative model will highly increase the training time of the model. (4) AutoIV [33] model proposes a framework for learning both instrumental factors and confounders using MI in representation learning, but ignores the learning of risk factors and high computational complexity of the conditional distributions in this algorithm. MIM-DRCFR [25] shares information in learning latent factors through a multi-task learning framework and ensures the independence of these factors by combining a classifier with restrict discrepancy distance and through MI regularization. Additionally, features in the shared underlying representation that are too abstract or generic may lead to insufficient information when disentangling factors. However, the two methods mainly focus on counterfactual prediction rather than causal effects estimation.

In summary, existing algorithms for estimating treatment effects: (1) Innocently handling non-confounders when balancing treatment variables on different treatment groups. (2) Inaccurately identifying confounders when disentangling treatment variables. (3) Difficulty in effectively reducing confounding bias by addressing confounders. (4) Existing models that are effectively disentangled are overly complex, dramatically increasing the models' training time. Therefore, we propose a DRL_{ECB} model to address the above difficulties.

3. Preliminaries

Table 1 shows notations and their descriptions in our model. We focus on treatment effect estimation from observational dataset $D = (x_i, t_i, y_i(t_i))_{i=1}^{N}$, where N is the number of data samples, x_i is the input treatment variables referring individual context characteristics, binary treatment $t_i \in \{0, 1\}$, $t_i = 0$ indicates the ith individual is control group and $t_i = 1$ indicates the individual is treatment group, $y_i(0)$ denote the observed outcome of i if it were controlled, $y_i(1)$ denote the observed outcome of i if it were treated. In the observational dataset, when $y_i(t_i)$ is the observed factual outcome, then $y_i(1 - t_i)$ is an unobserved outcome, and we cannot observe both of these outcomes at the same time.

Definition 1 (*Causal Effect [35]*)**.** Causal effect is when something happens or is happening based on something that occurred or is occurring. For example, Y happened due to X. The outcome of Y is strong or weak because of how well or how much X worked.

As the foundation of causal inference theory, The concept of causal effect was officially introduced by Donald Rubin in 1976 and further explained by Rubin and Paul Rosenbaum in 1983 as well as Paul Holland in 1986, resulting in the Rubin–Rosenbaum–Holland (RRH) theory of causal inference.

Definition 2 (*Disentangled Representation [36,37]*)**.** Disentangled representation refers to the separation of distinct, independent and informative generating factors from observational data to obtain a representation of underlying factors hidden in the observable data in the form of representations.

Definition 3 (*Confounding Bias [38]*)**.** When a variable causes a spurious correlation between variable T and the outcome Y, the variable is called a confounder factor/ variable of T and Y. The spurious correlation caused by confounder factors is considered a confounding bias.

Table 1
Summary of notations and descriptions.

Notations	Description
D	$(x_i, t_i, y_i(t_i))_{i=1^N}$ are observational data with N samples, $D = (x_i, t_i, y_i(t_i))_{i=1^N}$
X	Treatment variables
$\Gamma(X)$	Instrumental factors
$\Delta(X)$	Confounders
$Y(X)$	Risk factors
T	Treatment, $T \in \{0, 1\}$. $T = 0/1$ indicates the control /treatment groups, respectively.
Y^0	Outcome of control group
Y^1	Outcome of treatment group
x_i	Treatment variables of ith sample, $x_i \in X$
t_i	$t_i \in T$, $t_i = 0$ indicates the ith sample is controlled $t_i = 1$ indicates ith sample is treated
$y_i(t_i)$	Observed outcome of t_i
$y_i(1 - t_i)$	Unobserved outcome of $1 - t_i$
y_i	Observed outcome, $y_i \in \{y_i(0), y_i(1)\}$. $y_i(1)$, $y_i(0)$ denote the outcomes of treatment or control groups, respectively
$y_i'(0)$	Estimation outcome of $t_i = 0$
$y_i'(1)$	Estimation outcome of $t_i = 1$
θ	Variational approximation parameter
$q_\theta(\cdot\|\cdot)$	Variational approximation function
k_i	$k_i \in N$, k_i is randomly selected from $\{1, 2, \ldots, N\}$
$\log q_\theta(\cdot\|\cdot)$	Log-likelihood function
ITE	Individual/sample treatment effect
ATE	Average treatment effect
ϵ_{ATE}	Absolute error of ATE
MI	Mutual Information
$\Phi_\Gamma(X)$	Instrumental representation of $\Gamma(X)$
$\Phi_\Gamma^c(X)$	Instrumental representation with confounders
$\Phi_\Delta(X)$	Confounding representation of $\Delta(X)$
$\Phi_\Delta^i(X)$	Confounding representation with instrumental factors
$\Phi_Y(X)$	Risk representation of $Y(X)$
$\Phi_Y^c(X)$	Risk representation with confounders
$\Phi_\Delta'(X)$	Balanced $\Phi_\Delta(X)$
$\mathcal{L}(\theta)^{lld}$	Loss of the conditional log-likelihood
$\mathcal{L}(\theta)^{mi}$	MI bound loss
$\mathcal{L}_\Delta^{bal}$	Loss of balance confounding representation
$\mathcal{L}_{\Delta'Y}^{pred}$	Predictive loss
$\mathcal{L}_{total}$	Total losses, $\mathcal{L}_{total} = \mathcal{L}(\theta)^{lld} + \mathcal{L}(\theta)^{mi} + \mathcal{L}_\Delta^{bal} + \mathcal{L}_{\Delta'Y}^{pred}$

Definition 4 (*Mutual Information, MI [39]*)**.** MI is a basic measure of dependence between two random variables.

MI between variables $\boldsymbol{x}$ and $\boldsymbol{y}$ is defined as:

$$I(\boldsymbol{x}; \boldsymbol{y}) = \sum_{\boldsymbol{x},\boldsymbol{y}} p(\boldsymbol{x}, \boldsymbol{y}) \log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})} = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log \frac{p(\boldsymbol{x}, \boldsymbol{y})}{p(\boldsymbol{x})p(\boldsymbol{y})}]. \tag{1}$$

Moreover, we assume that the following three fundamental assumptions for estimating the treatment effect are satisfied [40]:

Assumption 1 (*Stable Unit Treatment Value, SUTV [34]*)**.** The potential outcomes for any unit do not vary with the treatment assigned to other units, and for each unit, there are no different forms or versions of each treatment level, which lead to different potential outcomes.

Assumption 2 (*Ignorability [25]*)**.** The treatment assignment mechanism is independent of the potential outcome when conditioning the treatment variables, Formally: $T \perp \left(Y^0, Y^1\right) | X$.

Assumption 3 (*Positivity [25]*)**.** Each unit has a non-zero probability of being assigned to each treatment when given the observed contexts, i.e., $0 < p(T = 1|X) < 1$.

4. DRL_{ECB} model

4.1. Problem setting

We propose a Disentangled Representation Learning model with Excluding Confounding Bias for causal effect estimation (DRL_{ECB}).

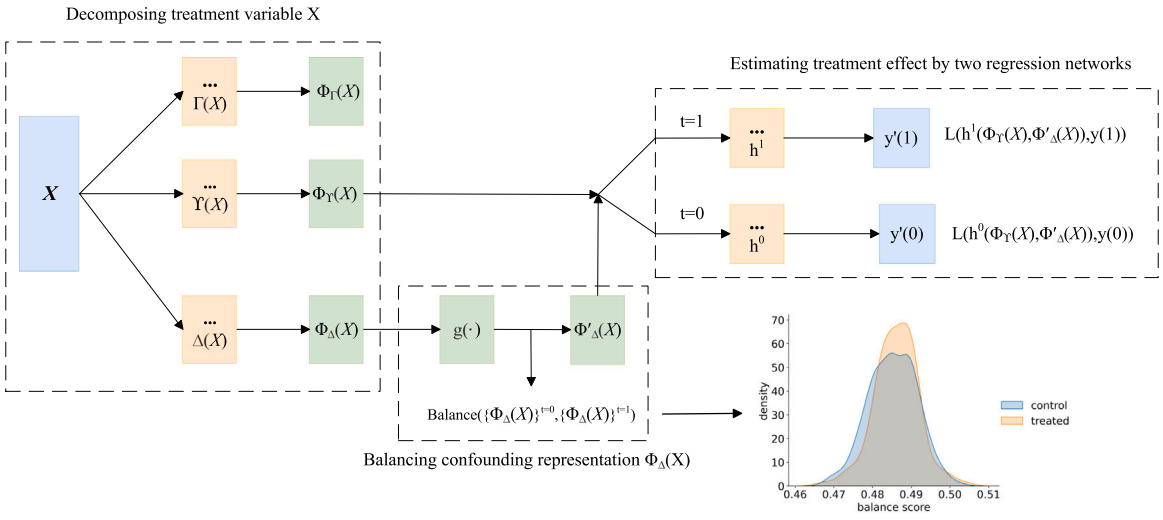


Fig. 3. The framework of our DRL_{ECB} model. The input observational data contains the treatment variables X, the binary treatment $t \in \{0 : control, 1 : treatment\}$ denotes the choice of the treatment, and $y'(0)$, $y'(1)$ denotes the outcomes obtained by applying the treatments mentioned above. X first learns the representation through a neural network and then obtains three mutually independent representations $\Phi_\Gamma(X), \Phi_\Delta(X), \Phi_Y(X)$ through MI decomposition, $g(\cdot)$ denotes the balance score, and $\text{Balance}(\{\Phi_\Delta(X)\}^{t=0}, \{\Phi_\Delta(X)\}^{t=1})$ realizes the confounding representation balance between treatment and control groups.

The framework of DRL_{ECB} includes the following three parts in detail (Fig. 3):

- Disentangling treatment variables X into three uncorrelated representations $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$ via MI, namely instrumental representation, confounding representation, and risk representation, respectively.
- Balancing confounding representation $\Phi_\Delta(X)$ to eliminate confounding bias by minimizing the squared difference between the reweighed treatment and control groups.
- Using balanced $\Phi_\Delta(X)$ and $\Phi_Y(X)$ to estimate treatment effects by regression networks $h^0(\Phi'_\Delta(X), \Phi_Y(X))$ and $h^1(\Phi'_\Delta(X), \Phi_Y(X))$, in which $\Phi'_\Delta(X)$ denotes balanced $\Phi_\Delta(X)$.

Inspired by Contrastive Log-ratio Upper Bound (CLUB) [39], which achieves inter-variable disentanglement by minimizing MI, we judge independence between variables by minimizing MI to mine disentangled representations. Let treatment variables X and Y be the model's input and outcome, and treatment $T = 1/0$ represents treatment/control groups. We represent three factors $\Gamma(X), \Delta(X), Y(X)$ by learned $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$ and $\Phi_Y(X)$, respectively. The objectives of our DRL_{ECB} are to learn disentangled representations of $\Gamma(X), \Delta(X), Y(X)$, from X and use them to estimate the causal effect of T and Y. $\Phi_\Gamma(X)$ is correlated with T and conditionally independent of Y given treatment T; $\Phi_\Delta(X)$ is correlated with T and Y; $\Phi_Y(X)$ is correlated with Y and uncorrelated with T. The basic settings of achieving DRL_{ECB} are generalized below. (1) Utilize sample pairs $(x_i, y_i)_{i=1}^N$ from distribution $p(y|x)$ and the variational distribution $q_\theta(y|x)$ with parameter θ to approximate $p(y|x)$. (2) Maximize/minimize the variational approximations $q_\theta(y_i|x_i)$ and $q_\theta(y_{k_i}|x_i)$ for the positive and negative sample pairs (x_i, y_i) and (x_i, y_{k_i}) to increase/decrease the correlation between x and y. ($k_i \in N$, randomly sampling k_i from N samples to reduce computational complexity speeds up computation.) (3) Estimate the balance score $g(\cdot)$ of $\Phi_\Delta(X)$, then use the score to reweight treatment variables X and minimize X's variance in treatment and control groups to obtain $\Phi'_\Delta(X)$. (4) Construct regression networks h^0 and h^1 on control ($t = 0$) and treatment group ($t = 1$), respectively, then use $\Phi'_\Delta(X)$ and $\Phi_Y(X)$ as input of h^0 and h^1 to estimate treatment effects $y'(0)$ and $y'(1)$.

4.2. Learning disentangled representation

4.2.1. Instrumental factors representation $\Phi_\Gamma(X)$

Due to learned $\Phi_\Gamma(X)$ comprise confounders $\Delta(X)$, which is correlated with T and Y, we first learn the representation of $\Gamma(X)$ with confounders $\Phi^c_\Gamma(X)$ before learning $\Phi_\Gamma(X)$.

Learning $\Phi^c_\Gamma(X)$. We use the variational distribution $q_{\theta_{\Gamma T}}(T|\Phi^c_\Gamma(X))$ with parameters $\theta_{\Gamma T}$ to approximate the true conditional distribution $\mathbb{P}(T|\Phi^c_\Gamma(X))$. The log-likelihood loss of $q_{\theta_{\Gamma T}}(T|\Phi^c_\Gamma(X))$ with N samples are below:

$$\mathcal{L}(\theta_{\Gamma T})^{lld} = -\frac{1}{N}\sum_{i=1}^{N} \log q_{\theta_{\Gamma T}}(\boldsymbol{t}_i|\phi^c_\Gamma(\boldsymbol{x}_i)). \tag{2}$$

We minimize Eq. (2) to get approximate distribution of $q_{\theta_{\Gamma T}}$, and then maximize MI between $\Phi^c_\Gamma(X)$ and T by Eq. (3) to achieve their correlation.

$$\begin{aligned}\mathcal{L}(\theta_{\Gamma T})^{mi} = -\frac{1}{N}\sum_{i=1}^{N}(&\log q_{\theta_{\Gamma T}}(\boldsymbol{t}_i|\phi^c_\Gamma(\boldsymbol{x}_i)) \\ &- \log q_{\theta_{\Gamma T}}(\boldsymbol{t}_{k_i}|\phi^c_\Gamma(\boldsymbol{x}_i))),\end{aligned} \tag{3}$$

in which $\log q_{\theta_{\Gamma T}}(\boldsymbol{t}_i|\phi^c_\Gamma(\boldsymbol{x}_i))$ and $\log q_{\theta_{\Gamma T}}(\boldsymbol{t}_{k_i}|\phi^c_\Gamma(\boldsymbol{x}_i))$ denote the conditional log-likelihood of positive and negative sample pair $(\phi^c_\Gamma(\boldsymbol{x}_i), \boldsymbol{t}_i)$ and $(\phi^c_\Gamma(\boldsymbol{x}_i), \boldsymbol{t}_{k_i})$ respectively, where $k_i \in N$, N is sample numbers. To get $\Phi^c_\Gamma(X)$, we minimize Eq. (3) by maximizing differences between $(\phi^c_\Gamma(\boldsymbol{x}_i), \boldsymbol{t}_i)$ and $(\phi^c_\Gamma(\boldsymbol{x}_i), \boldsymbol{t}_{k_i})$.

Learning $\Phi_\Gamma(X)$. To get $\Phi_\Gamma(X)$, which is conditionally independent of Y given treatment T, we intend to remove confounders from $\Phi^c_\Gamma(X)$. We first use variational distribution $q_{\theta_{\Gamma Y}}(Y|\Phi^c_\Gamma(X))$ with parameters $\theta_{\Gamma Y}$ to approximate the true conditional distribution $\mathbb{P}(Y|\Phi^c_\Gamma(X))$. The log-likelihood loss of $q_{\theta_{\Gamma Y}}(Y|\Phi^c_\Gamma(X))$ are below:

$$\mathcal{L}(\theta_{\Gamma Y})^{lld} = -\frac{1}{N}\sum_{i=1}^{N} \log q_{\theta_{\Gamma Y}}(\boldsymbol{y}_i|\phi^c_\Gamma(\boldsymbol{x}_i)). \tag{4}$$

Then, we remove confounders from $\Phi^c_\Gamma(X)$ by conditional independence (d-separation) to get $\Phi_\Gamma(X)$ and make $\Phi_\Gamma(X) \perp Y \mid T$, (note: $(\Phi_\Gamma(X))$) is $(\Phi^c_\Gamma(X)$ without confounders). Given T, the impact of $\Phi^c_\Gamma(X)$ on Y can be blocked. Specially, we exploit minimizing Eq. (4) to get variational approximation $q_{\theta_{\Gamma Y}}$ and then minimize MI between

$\Phi_\Gamma^c(X)$ and Y to achieve independent between $\Phi_\Gamma(X)$ and Y given T as follows:

$$\mathcal{L}(\theta_{\Gamma Y})^{mi} = \frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Gamma Y}}(\boldsymbol{y}_i|\phi_\Gamma^c(\boldsymbol{x}_i)) - \log q_{\theta_{\Gamma Y}}(\boldsymbol{y}_{k_i}|\phi_\Gamma^c(\boldsymbol{x}_i))), \tag{5}$$

in which $\log q_{\theta_{\Gamma Y}}(\boldsymbol{y}_i|\phi_\Gamma^c(\boldsymbol{x}_i))$ and $\log q_{\theta_{\Gamma Y}}(\boldsymbol{y}_{k_i}|\phi_\Gamma^c(\boldsymbol{x}_i))$ represent respectively the conditional log-likelihood of positive and negative sample pair, $(\phi_\Gamma^c(\boldsymbol{x}_i), \boldsymbol{y}_i)$ and $(\phi_\Gamma^c(\boldsymbol{x}_i), \boldsymbol{y}_{k_i})$. We obtain $\Phi_\Gamma(X)$ by minimizing Eq. (5) (i.e., minimizing the difference between positive and negative samples) to discard confounders from $\Phi_\Gamma^c(X)$ and get $\Phi_\Gamma(X)$. That is to say, given T and Y are known, the confounders in $\Phi_\Gamma^c(X)$ are removed. Significantly, removed confounders are relevant with T and Y due to learned representations $\Phi_\Gamma(X)$ being conditional on T and Y.

4.2.2. Confounders representation $\Phi_\Delta(X)$

Considering $\Phi_\Delta(X)$ is correlated with T and Y simultaneously. The instrumental factors $\Gamma(X)$ are also correlated with T; we need to discard the effect of $\Gamma(X)$ from $\Phi_\Delta(X)$. $\Phi_\Delta^i(X)$ represent confounding representation with $\Gamma(X)$.

Learning $\Phi_\Delta^i(X)$. We first use variational distribution $q_{\theta_{\Delta T}}(T|\Phi_\Delta^i(X))$ with parameters $\theta_{\Delta T}$ to approximate conditional distribution $\mathbb{P}(T|\Phi_\Delta^i(X))$. The log-likelihood loss of $q_{\theta_{\Delta T}}(T|\Phi_\Delta^i(X))$ are below:

$$\mathcal{L}(\theta_{\Delta T})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{\Delta T}}(\boldsymbol{t}_i|\phi_\Delta^i(\boldsymbol{x}_i)). \tag{6}$$

By minimizing Eq. (6), we can obtain approximation $q_{\theta_{\Delta T}}(T|\Phi_\Delta^i(X))$, and then maximize MI between $\Phi_\Delta^i(X)$ and T to achieve correlation between them as follows:

$$\mathcal{L}(\theta_{\Delta T})^{mi} = -\frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Delta T}}(\boldsymbol{t}_i|\phi_\Delta^i(\boldsymbol{x}_i)) - \log q_{\theta_{\Delta T}}(\boldsymbol{t}_{k_i}|\phi_\Delta^i(\boldsymbol{x}_i))). \tag{7}$$

Further, we minimize Eq. (7) to get $\Phi_\Delta^i(X)$ via maximizing differences between the positive $(\phi_\Delta^i(\boldsymbol{x}_i), \boldsymbol{t}_i)$ and negative $(\phi_\Delta^i(\boldsymbol{x}_i), \boldsymbol{t}_{k_i})$ sample pairs.

Learning $\Phi_\Delta(X)$. To remove instrumental factors from $\Phi_\Delta^i(X)$, we intent to make $\Phi_\Delta^i(X)$ is correlated with Y to get $\Phi_\Delta(X)$. Specifically, we utilize variational distribution $q_{\theta_{\Delta Y}}(Y|\Phi_\Delta^i(X))$ to approximate conditional distribution $\mathbb{P}(Y|\Phi_\Delta^i(X))$. The log-likelihood loss of $q_{\theta_{\Delta Y}}(Y|\Phi_\Delta^i(X))$ as follows:

$$\mathcal{L}(\theta_{\Delta Y})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{\Delta Y}}(\boldsymbol{y}_i|\phi_\Delta^i(\boldsymbol{x}_i)). \tag{8}$$

Then, we minimize Eq. (8) to get $q_{\theta_{\Delta Y}}$ and maximize MI between $\Phi_\Delta^i(X)$ and Y by minimizing Eq. (9) to get $\Phi_\Delta(X)$.

$$\mathcal{L}(\theta_{\Delta Y})^{mi} = -\frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Delta Y}}(\boldsymbol{y}_i|\phi_\Delta^i(\boldsymbol{x}_i)) - \log q_{\theta_{\Delta Y}}(\boldsymbol{y}_{k_i}|\phi_\Delta^i(\boldsymbol{x}_i))). \tag{9}$$

4.2.3. Risk factors representation $\Phi_Y(X)$

Due to both $\Phi_Y(X)$ and $\Delta(X)$ are correlated with Y, we learn $\Phi_Y(X)$ with $\Delta(X)$ before learning $\Phi_Y(X)$.

Learning $\Phi_Y^c(X)$. We use the variational distribution $q_{\theta_{YY}}(Y|\Phi_Y^c(X))$ with parameters θ_{YY} to approximate conditional distribution $\mathbb{P}(Y|\Phi_Y^c(X))$. The log-likelihood loss of $q_{\theta_{YY}}(Y|\Phi_Y^c(X))$ are below:

$$\mathcal{L}(\theta_{YY})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{YY}}(\boldsymbol{y}_i|\phi_Y^c(\boldsymbol{x}_i)). \tag{10}$$

We minimize Eq. (10) to get $q_{\theta_{YY}}$ with parameters θ_{YY}, we maximize MI between $\Phi_Y^c(X)$ and Y to achieve correlation between them as follows:

$$\mathcal{L}(\theta_{YY})^{mi} = -\frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{YY}}(\boldsymbol{y}_i|\phi_Y^c(\boldsymbol{x}_i)) - \log q_{\theta_{YY}}(\boldsymbol{y}_{k_i}|\phi_Y^c(\boldsymbol{x}_i))). \tag{11}$$

We minimize Eq. (11) to obtain $\Phi_Y^c(X)$ by maximizing the difference between positive $(\phi_Y^c(\boldsymbol{x}_i), \boldsymbol{y}_i)$ and negative $(\phi_Y^c(\boldsymbol{x}_i), \boldsymbol{y}_{k_i})$ samples pairs from $\log q_{\theta_{YY}}(\boldsymbol{y}_i|\phi_Y^c(\boldsymbol{x}_i))$ and $\log q_{\theta_{YY}}(\boldsymbol{y}_{k_i}|\phi_Y^c(\boldsymbol{x}_i))$, respectively.

Learning $\Phi_Y(X)$. To get $\Phi_Y(X)$, we remove $\Delta(X)$ in $\Phi_Y^c(X)$ by making $\Phi_Y^c(X)$ uncorrelated of T.

We approximate the true conditional distribution $\mathbb{P}(T|\Phi_Y^c(X))$ by $q_{\theta_{YT}}(T|\Phi_Y^c(X))$ with parameter θ_{YT} as below:

$$\mathcal{L}(\theta_{YT})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{YT}}(\boldsymbol{t}_i|\phi_Y^c(\boldsymbol{x}_i)). \tag{12}$$

Then, minimize Eq. (12) to get $q_{\theta_{YT}}(T|\Phi_Y^c(X))$, and then minimize MI of $\Phi_Y^c(X)$ and T to achieve uncorrelation between them as follows:

$$\mathcal{L}(\theta_{YT})^{mi} = \frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{YT}}(\boldsymbol{t}_i|\phi_Y^c(\boldsymbol{x}_i)) - \log q_{\theta_{YT}}(\boldsymbol{t}_{k_i}|\phi_Y^c(\boldsymbol{x}_i))). \tag{13}$$

Then, we minimize the difference between positive $(\phi_Y^c(\boldsymbol{x}_i), \boldsymbol{t}_i)$ and negative $(\phi_Y^c(\boldsymbol{x}_i), \boldsymbol{t}_{k_i})$ sample pairs to get $\Phi_Y(X)$ by minimizing Eq. (13).

4.3. Regularizing $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$

Considering learned representations $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$ are conditional on T and Y, This means that these representations may contain mutual information related to T and Y, which can constrain the generalization of models in complex scenarios, e.g., nonlinear high-dimensional spaces. Therefore, we design regularization items by minimizing MI between $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$, which can enhance the independence of learned representations, to improve generalizations.

Minimizing MI between $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$. We exploit variational distribution $q_{\theta_{\Gamma\Delta}}(\Phi_\Delta(X)|\Phi_\Gamma(X))$ to approximate the conditional distribution $\mathbb{P}(\Phi_\Delta(X)|\Phi_\Gamma(X))$ for realizing the independence of $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$.

$$\mathcal{L}(\theta_{\Gamma\Delta})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{\Gamma\Delta}}(\phi_\Delta(\boldsymbol{x}_i)|\phi_\Gamma(\boldsymbol{x}_i)), \tag{14}$$

$$\mathcal{L}(\theta_{\Gamma\Delta})^{mi} = \frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Gamma\Delta}}(\phi_\Delta(\boldsymbol{x}_i)|\phi_\Gamma(\boldsymbol{x}_i)) - \log q_{\theta_{\Gamma\Delta}}(\phi_\Delta(\boldsymbol{x}_{k_i})|\phi_\Gamma(\boldsymbol{x}_i))). \tag{15}$$

We minimize Eq. (14) to obtain an variational approximation of $q_{\theta_{\Gamma\Delta}}(\Phi_\Delta(X)|\Phi_\Gamma(X))$ and minimizing Eq. (15) to regularize the $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$.

Minimizing MI between $\Phi_\Gamma(X)$ and $\Phi_Y(X)$. We minimize MI between the $\Phi_\Gamma(X)$ and $\Phi_Y(X)$ to achieve mutual independence of the two representations.

$$\mathcal{L}(\theta_{\Gamma Y})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{\Gamma Y}}(\phi_Y(\boldsymbol{x}_i)|\phi_\Gamma(\boldsymbol{x}_i)), \tag{16}$$

$$\mathcal{L}(\theta_{\Gamma Y})^{mi} = \frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Gamma Y}}(\phi_Y(\boldsymbol{x}_i)|\phi_\Gamma(\boldsymbol{x}_i)) - \log q_{\theta_{\Gamma Y}}(\phi_Y(\boldsymbol{x}_{k_i})|\phi_\Gamma(\boldsymbol{x}_i))). \tag{17}$$

We minimize Eq. (16) to obtain variational approximation $q_{\theta_{\Gamma Y}}(\Phi_Y(X)|\Phi_\Gamma(X))$ and minimize MI between $\Phi_\Gamma(X)$ and $\Phi_Y(X)$ by minimizing Eq. (17).

Minimizing MI between $\Phi_\Delta(X)$ and $\Phi_Y(X)$. We minimize MI between $\Phi_\Delta(X)$ and $\Phi_Y(X)$ to achieve independence of the two representations.

$$\mathcal{L}(\theta_{\Delta Y})^{lld} = -\frac{1}{N}\sum_{i=1}^{N}\log q_{\theta_{\Delta Y}}(\phi_\Delta(\boldsymbol{x}_i)|\phi_Y(\boldsymbol{x}_i)), \tag{18}$$

$$\mathcal{L}(\theta_{\Delta Y})^{mi} = \frac{1}{N}\sum_{i=1}^{N}(\log q_{\theta_{\Delta Y}}(\phi_\Delta(\boldsymbol{x}_i)|\phi_Y(\boldsymbol{x}_i)) - \log q_{\theta_{\Delta Y}}(\phi_\Delta(\boldsymbol{x}_{k_i})|\phi_Y(\boldsymbol{x}_i))). \tag{19}$$

Similar to the above, we minimize Eq. (18) to obtain variational approximation $q_{\theta_{\Delta Y}}(\Phi_\Delta(X)|\Phi_Y(X))$ for the conditional distribution $\mathbb{P}(\Phi_\Delta(X)|\Phi_Y(X))$ and regularize $\Phi_\Delta(X)$ and $\Phi_Y(X)$ via minimizing Eq. (19).

4.4. Balancing confounding representation $\Phi_\Delta(X)$

$\Phi_\Delta(X)$ causes the distributional differences between treatment and control groups, which can result in confounding bias. Therefore, we intend to eliminate the bias by balancing $\Phi_\Delta(X)$ two groups.

Given $\Phi_\Delta(X)$, we get $g(\Phi_\Delta(X))$ by sigmoid function. Then, we utilize $\frac{1}{1-g(\Phi_\Delta(X))}$ and $\frac{1}{g(\Phi_\Delta(X))}$ to reweight the treatment and control groups, respectively, and the weight loss is as follows:

$$\mathcal{L}_\Delta^{bal} = \frac{1}{d}\sum_{k=1}^{d}\left(\frac{\sum_{i=1}^{N}\frac{x_{i,k}\cdot t_i}{g(\phi_\Delta(x_i))}}{\sum_{i=1}^{N}\frac{t_i}{g(\phi_\Delta(x_i))}} - \frac{\sum_{i=1}^{N}\frac{x_{i,k}\cdot(1-t_i)}{1-g(\phi_\Delta(x_i))}}{\sum_{i=1}^{N}\frac{1-t_i}{1-g(\phi_\Delta(x_i))}}\right)^2. \tag{20}$$

In Eq. (20), d denotes the sum of treatment variables, where $x_{i,k}$ is the kth treatment variable of sample i. Then, we minimize the squared deviation between the treatment and control groups by minimizing Eq. (20) to make the confounding representation achieve similar distributions.

$$\frac{\sum_{i=1}^{N}\frac{x_{i,k}\cdot t_i}{g(\phi_\Delta(x_i))}}{\sum_{i=1}^{N}\frac{t_i}{g(\phi_\Delta(x_i))}} \approx \frac{\sum_{i=1}^{N}\frac{x_{i,k}\cdot(1-t_i)}{1-g(\phi_\Delta(x_i))}}{\sum_{i=1}^{N}\frac{1-t_i}{1-g(\phi_\Delta(x_i))}} \tag{21}$$

By minimizing Eq. (20), we get Eq. (21) and the values of $g(\phi_\Delta(\boldsymbol{x}_i))$ and $1 - g(\phi_\Delta(\boldsymbol{x}_i))$ are constantly similar and approaches 0.5 which generate balanced $\Phi_\Delta(X)$, i.e., $\Phi'_\Delta(X)$.

4.5. Estimating causal effects

Inspired by [27,41,42], we first train two regression networks h^0 and h^1 to get predicated outcomes of control and treatment groups, respectively. Given $\Phi'_\Delta(X)$ and $\Phi_Y(X)$ as inputs, and we can obtain the loss $\mathcal{L}_{\Delta'Y}^{pred}$ between real and prediction results. Let $h^{t_i}(\cdot)$ denote the function learned by h^0 and h^1, $t_i \in \{0,1\}$. This loss function is as follows:

$$\mathcal{L}_{\Delta'Y}^{pred} = \mathcal{L}[y_i, h^{t_i}(\Phi'_\Delta(X), \Phi_Y(X))], \tag{22}$$

where y_i is the observed outcome. We minimize the mean squared error (MSE) to get the predicted loss $\mathcal{L}_{\Delta'Y}^{pred}$ in Eq. (22).

To estimate treatment effects, we minimize Eqs. (2), (4), (6), (8), (10), (12), (14), (16) and (18) to optimize the parameters $\theta_{\Gamma T}$, $\theta_{\Gamma Y}$, $\theta_{\Delta T}$, $\theta_{\Delta Y}$, θ_{YY}, θ_{YT}, $\theta_{\Gamma\Delta}$, $\theta_{\Gamma Y}$, and $\theta_{\Delta Y}$, respectively, with each variational distribution approximating the corresponding conditional distribution. We combine all the losses to obtain the final conditional log-likelihood as follows:

$$\begin{aligned}\mathcal{L}^{lld} &= \mathcal{L}(\theta_{\Gamma T})^{lld} + \mathcal{L}(\theta_{\Gamma Y})^{lld} + \mathcal{L}(\theta_{\Delta T})^{lld} \\ &+ \mathcal{L}(\theta_{\Delta Y})^{lld} + \mathcal{L}(\theta_{YY})^{lld} + \mathcal{L}(\theta_{YT})^{lld} \\ &+ \mathcal{L}(\theta_{\Gamma\Delta})^{lld} + \mathcal{L}(\theta_{\Gamma Y})^{lld} + \mathcal{L}(\theta_{\Delta Y})^{lld}.\end{aligned} \tag{23}$$

Note that each term in Eq. (23) independently optimizes the appropriate parameters. Further, we combine all MI loss of Eqs. (3), (5), (7), (9), (11), (13), (15), (17) and (19) as follows:

$$\begin{aligned}\mathcal{L}^{mi} &= \mathcal{L}(\theta_{\Gamma T})^{mi} + \mathcal{L}(\theta_{\Gamma Y})^{mi} + \mathcal{L}(\theta_{\Delta T})^{mi} \\ &+ \mathcal{L}(\theta_{\Delta Y})^{mi} + \mathcal{L}(\theta_{YY})^{mi} + \mathcal{L}(\theta_{YT})^{mi} \\ &+ \mathcal{L}(\theta_{\Gamma\Delta})^{mi} + \mathcal{L}(\theta_{\Gamma Y})^{mi} + \mathcal{L}(\theta_{\Delta Y})^{mi}.\end{aligned} \tag{24}$$

Building upon it and taking the above losses into account, we propose the following objective total loss $\mathcal{L}_{total}$ for causal effect estimation below:

$$\mathcal{L}_{total} = \mathcal{L}^{lld} + \mathcal{L}^{mi} + \beta\cdot\mathcal{L}_\Delta^{bal} + \mathcal{L}_{\Delta'Y}^{pred}, \tag{25}$$

where $\mathcal{L}^{lld}$ is all the losses of variational approximation; $\mathcal{L}^{mi}$ is all the MI constraints loss functions; $\mathcal{L}_\Delta^{bal}$ is balancing loss between the treatment and control groups; $\mathcal{L}_{\Delta'Y}^{pred}$ is the prediction loss for observed outcomes. β controls the relative importance of $\mathcal{L}_\Delta^{bal}$ in the overall loss $\mathcal{L}_{total}$.

Algorithm 1 DRL_{ECB} algorithm

1: **Input:** Observational data $(x_i, t_i, y_i(t_i))_{i=1}^N$.
Initialize variational distribution parameters $\theta_{\Gamma T}$, $\theta_{\Gamma Y}$, $\theta_{\Delta T}$, $\theta_{\Delta Y}$, θ_{YT}, θ_{YY}, $\theta_{\Gamma\Delta}$, $\theta_{\Gamma Y}$, $\theta_{\Delta Y}$; Hyperparameter β=1.0.
2: **Loss function:** $\mathcal{L}^{lld}$, $\mathcal{L}^{mi}$, $\mathcal{L}_\Delta^{bal}$ and $\mathcal{L}_{\Delta'Y}^{pred}$.
3: **Components:** Three representation learning networks $\Phi_\Gamma(X), \Phi_\Delta(X), \Phi_Y(X)$, balanced confounding representation $\Phi'_\Delta(X)$, two regression networks h^0 and h^1 for the predicted outcomes.
4: **for** $i = 0, 1, 2, ..., N$ **do**
5: Update parameters $\theta_{\Gamma T}$, $\theta_{\Gamma Y}$, $\theta_{\Delta T}$, $\theta_{\Delta Y}$, θ_{YT}, θ_{YY}, $\theta_{\Gamma\Delta}$, $\theta_{\Gamma Y}$, $\theta_{\Delta Y}$, by minimizing $\mathcal{L}^{lld}$ in Eq. (23).
6: Update $\Phi_\Gamma(X), \Phi_\Delta(X), \Phi_Y(X)$ representations by minimizing $\mathcal{L}^{mi}$ in Eq. (24).
7: Update confounding balancing loss $\mathcal{L}_\Delta^{bal}$ by minimizing Eq. (20) to get $\Phi'_\Delta(X)$.
8: Update predicted loss $\mathcal{L}_{\Delta'Y}^{pred}$ in two regression networks by minimizing Eq. (22).
9: **Output:** $y'(0), y'(1)$.
10: **end for**

4.6. DRL_{ECB} algorithm

We design our DRL_{ECB} algorithm based on the above analyses. Furthermore, the pseudocode of DRL_{ECB} is presented as Algorithm 1.

DRL_{ECB} consists of three steps to train the causal effect estimation model, including (1) data preprocessing and parameters initialization, (2) calculating the loss and updating the parameter (lines:4–8), and (3) outputting the prediction results (line 9). When a batch of training instances arrives, the model will process them one by one according to lines 1–10. Specifically, DRL_{ECB} initially obtains three independent representations by minimizing the log-likelihood function $\mathcal{L}^{lld}$ and the MI loss $\mathcal{L}^{mi}$ (lines:5–6); then implements the balance of the confounding representation in the treatment and control groups to eliminate the confounding bias (line 7); and finally, $\Phi'_\Delta(X)$ and $\Phi_Y(X)$ are used as inputs to implement the treatment effect estimation using the regression network (lines:8–9).

To effectively realize the disentangled representation, DeR-CFR [21] and TEDVAE [32] use orthogonal constraints and VAE respectively, both of which greatly increase the model complexity and lead to a significant increase in the computational complexity; AutoIV [33] and our DRL_{ECB} use MI to realize the decomposition. The complexity of computing MI minimization loss in AutoIV is $O(N^2)$ due to Eq.13 in [33]. Computing MI losses for training disentangling representation in DRL_{ECB} can be reduced to $O(N)$ because of Eq. (24) and adopted random negative sampling to achieve an unbiased estimation.

5. Theoretical guarantees

In this section, we theoretically analyze the rationality of MI minimization to achieve disentangling in DRL_{ECB}. Inspired by CLUB [39], since the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ is unknown in our task, we approximate the conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ by using the variational distribution $q_\theta(\boldsymbol{y}|\boldsymbol{x})$ with parameters θ. The upper bound of MI between $\boldsymbol{x}$ and $\boldsymbol{y}$, termed $I_{vCLUB}(\boldsymbol{x};\boldsymbol{y})$, is defined as:

$$I_{vCLUB}(\boldsymbol{x};\boldsymbol{y}) = \mathbb{E}_{p(\boldsymbol{x},\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})] - \mathbb{E}_{p(\boldsymbol{x})}\mathbb{E}_{p(\boldsymbol{y})}[\log q_\theta(\boldsymbol{y}|\boldsymbol{x})]. \tag{26}$$

Building upon it, the following Theorems 1 and 2 mainly express the rationality of disentangling treatment variables by MI minimization. Theorem 3 aims to demonstrate that negative sampling can reduce computational complexity without compromising model performance.

Theorem 1. *For two random variables x and y, if the variational joint distribution $q_\theta(x,y) = q_\theta(x|y)p(x)$ satisfies the following inequality:*

$$KL(p(x,y) \parallel q_\theta(x,y)) \le KL(p(x)p(y) \parallel q_\theta(x,y)), \quad (27)$$

then $I(x;y) \le I_{vCLUB}(x;y)$ holds.

Proof. We calculate the gap Δ between $I_{vCLUB}(x;y)$ and $I(x;y)$:

$$\begin{aligned}\Delta &= I_{vCLUB}(x;y) - I(x;y)\\ &= \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)] - \mathbb{E}_{p(x)}\mathbb{E}_{p(y)}[\log q_\theta(y|x)]\\ &\quad -\mathbb{E}_{p(x,y)}[\log p(y|x) - \log p(y)]\\ &= \mathbb{E}_{p(x)p(y)}[\log \frac{p(y)}{q_\theta(y|x)}] - \mathbb{E}_{p(x,y)}[\log \frac{p(y|x)}{q_\theta(y|x)}]\\ &= \mathbb{E}_{p(x)p(y)}[\log \frac{p(x)p(y)}{q_\theta(x,y)}] - \mathbb{E}_{p(x,y)}[\log \frac{p(x,y)}{q_\theta(x,y)}]\\ &= KL(p(x)p(y) \parallel q_\theta(x,y)) - KL(p(x,y) \parallel q_\theta(x,y)).\end{aligned} \quad (28)$$

According to Eq. (27), $\Delta > 0$ Therefore, $I_{vCLUB}(x;y)$ is an upper bound on $I(x;y)$. If and only if x and y are independent, the $I(x;y) = I_{vCLUB}(x;y)$. Theorem 1 holds.

Theorem 2. *Suppose Eq. (27) in Theorem 1 is satisfied, then optimizing the variational distribution $q_\theta(y|x)$ by maximizing the log-likelihood function is equivalent to minimizing $KL(p(x,y) \parallel q_\theta(x,y))$.*

Proof. KL divergence minimization:

$$\begin{aligned}&min_\theta KL(p(x,y) \parallel q_\theta(x,y))\\ &\quad = min_\theta \mathbb{E}_{p(x,y)}[\log p(y|x)p(x) - \log q_\theta(y|x)p(x)]\\ &\quad = min_\theta \mathbb{E}_{p(x,y)}[\log p(y|x)] - \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)]\\ &\quad = min_\theta KL(p(y|x) \parallel q_\theta(y|x))\\ &\quad = max_\theta \mathbb{E}_{p(x,y)}[\log q_\theta(y|x)].\end{aligned} \quad (29)$$

Theorem 2 holds.

For given sample pairs $(x_i, y_i)_{i=1}^N$, the maximizing the log-likelihood function $\frac{1}{N}\sum_{i=1}^N \log q_\theta(y_i|x_i)$ is equivalent to $\mathbb{E}_{p(x,y)}[\log q_\theta(y|x)]$. According to Eq. (29), $\frac{1}{N}\sum_{i=1}^N \log q_\theta(y_i|x_i)$ is equivalent to $KL(p(x,y) \parallel q_\theta(x,y))$. For the sample pairs given above, $I_{vCLUB}(x;y)$ in Eq. (26) can be written as:

$$\hat{I}_{vCLUB}(x;y) = \frac{1}{N}\sum_{i=1}^N [\log q_\theta(y_i|x_i) - \frac{1}{N}\sum_{j=1}^N \log q_\theta(y_j|x_i)]. \quad (30)$$

In summary, we obtain a good variational approximation $q_\theta(y|x)$ by maximizing the log-likelihood function $\frac{1}{N}\sum_{i=1}^N \log q_\theta(y_i|x_i)$, and according to Theorem 1, $\hat{I}_{vCLUB}(x;y)$ is still an upper bound on MI ($I(x;y)$), then by minimizing the loss of $\hat{I}_{vCLUB}(x;y)$, we can reduce the correlation between the two variables and obtain independent x and y.

Theorem 3. *The log probability $\log q_\theta(y_{k_i}|x_i)$ of a random negative sample pair (x_i, y_{k_i}) ($k_i \in N$, randomly sampling k_i from N samples), instead of calculating the probabilities average of negative sample pairs $\frac{1}{N}\sum_{j=1}^N \log q_\theta(y_j|x_i)$ in Eq. (30), can achieve unbiased estimation and reduce computational complexity. The substituted formula is as follows:*

$$\hat{I}_{vCLUB-s}(x;y) = \frac{1}{N}\sum_{i=1}^N [\log q_\theta(y_i|x_i) - \log q_\theta(y_{k_i}|x_i)]. \quad (31)$$

Proof. The mathematical expectation $\mathbb{E}[\hat{I}_{vCLUB-s}(x;y)]$ of $\hat{I}_{vCLUB-s}(x;y)$ is as follows:

$$\begin{aligned}&\mathbb{E}[\hat{I}_{vCLUB-s}(x;y)]\\ &\quad = \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N [\log q_\theta(y_i|x_i) - \log q_\theta(y_{k_i}|x_i)]\right]\\ &\quad = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\log q_\theta(y_i|x_i)] - \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\log q_\theta(y_{k_i}|x_i)]\end{aligned} \quad (32)$$

Table 2
Summary of the real and synthetic datasets.

Dataset category	Datasets	Treatment variables	Treated/Controlled	Samples	Metrics
Real	IHDP	25	2/8	74 700	$PEHE$, ϵ_{ATE}
	Jobs	17	1/9	32 120	R_{pol}, ϵ_{ATT}
	Twins	30	5/5	11 400	$PEHE$, ϵ_{ATE}
Synthetic	$m_\Gamma\_m_\Delta\_m_Y$	$\{32 \sim 50\}$	5/5	3000	$PEHE$, ϵ_{ATE}

Notice that in Eq. (31), $k_i \in N$, randomly sampling k_i from N samples, thus $\mathbb{E}[\log q_\theta(y_{k_i}|x_i)] = \frac{1}{N}\sum_{j=1}^N \mathbb{E}[\log q_\theta(y_j|x_i)]$, so Eq. (32) can be written as:

$$\begin{aligned}&\mathbb{E}[\hat{I}_{vCLUB-s}(x;y)]\\ &\quad = \frac{1}{N}\sum_{i=1}^N \mathbb{E}[\log q_\theta(y_i|x_i)] - \frac{1}{N^2}\sum_{i=1}^N\sum_{j=1}^N \mathbb{E}[\log q_\theta(y_j|x_i)]\\ &\quad = \frac{1}{N}\sum_{i=1}^N [\log q_\theta(y_i|x_i) - \frac{1}{N}\sum_{j=1}^N \log q_\theta(y_j|x_i)]\\ &\quad = \mathbb{E}[\hat{I}_{vCLUB}(x;y)]\end{aligned} \quad (33)$$

According to Eq. (33), $\mathbb{E}[\hat{I}_{vCLUB-s}(x;y)] = \mathbb{E}[\hat{I}_{vCLUB}(x;y)]$ can achieve unbiased estimation and does not affect the accuracy of causal effect estimation. The computational complexity in each iteration is also reduced from $O(N^2)$ in Eq. (30) to $O(N)$ in Eq. (31). Theorem 3 holds.

6. Experiments

To cover the more general and challenging cases of DRL_{ECB} in the estimation of causal effects, we aim to answer the following research questions (RQs) in experiments.

- RQ1. Is our DRL_{ECB} superior to its rivals on treatment effect estimation algorithms?
- RQ2. (1) Is X are effectively disentangled by MI into $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$? (2) Is regularizing $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$ effective?
- RQ3. Is balanced confounding representation $\Phi'_\Delta(X)$ effective in eliminating confounding bias?

6.1. General settings

Datasets. Following previous research, we validate DRL_{ECB} on three real benchmark datasets: IHDP, Jobs, Twins, and eight synthetic datasets. The details of datasets are shown in Table 2.

- **Real datasets: IHDP** The Infant Health and Development Program (IHDP) conducted an RCT studying the effect of home visits on infant's future cognitive test scores. Hill [43] induced selection bias by removing a non-randomized subset of the treatment population from this RCT data to create a real observational dataset containing 747 units (139 treatment units, 608 control units) and 25 treatment variables related to the characteristics of the infants and their mothers. We experiment on 100 original split subsets with 90% for training and 10% for testing [21,28,32]
- **Real datasets: Jobs** LaLonde [44] combines an RCT (based on the National Supported Work program) with observational data to construct a binary categorical task called "Jobs" designed to estimate the effect of job training programs on employment status. The dataset includes 17 treatment variables: age, education level, and previous earnings. We conduct the experiment using the

Table 3
Summary of compared algorithms.

Algorithms	Descriptions	Parameters			
		Learning rate	Weight decay rate	Activation	Optimizer
CFR-MMD [27]	A classical causal effect estimation model based on MMD to balance the representation distribution.	0.001	0.0001	ReLU	Adam
CFR-WASS [27]	A classical causal effect estimation model based on WASS to balance the representation distribution.	0.001	0.0001	ReLU	Adam
CFR-ISW [28]	Based on propensity scores to weight the optimization function and eliminate bias to estimate causal effect.	0.001	0.001	ELU	Adam
SITE [29]	Utilizing MPDM and PDDM preserve local similarity information in the representation space while balancing data distribution.	0.001	0.001	ELU	Adam
DRRL [30]	Entropy balance method learns weights to minimize JS divergence between different treatment groups, and uses DR to estimate treatment effect.	0.05	\	ReLU	RMSProp
DR-CFR [31]	Minimize distribution discrepancy through MMD to decompose X for the estimation of causal effects.	0.001	0.0001	ReLU	Adam
TEDVAE [32]	Based on the VAE generative disentangle model realizes to disentangle X for estimation the treatment effect.	0.001	0.0001	ELU	ClippedAdam
DeR-CFR [21]	Using regression model, MMD, and orthogonal regularize to decompose X and eliminate bias for estimation the treatment effect.	0.001	1.0	ELU	Adam
MIM-DRCFR [25]	Learning shared features by sharing underlying representations then learning disentangled factors with factor-specific correlated networks and MI minimization.	0.001	\	ReLU	Adam
AutoIV [33]	Learning $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$ via MI maximization and minimization constraints, and achieving counterfactual predictions based on $\Phi_\Gamma(X)$.	0.001	0.00001	ReLU	Adam

LaLonde [27] experimental sample (297 treatment, 425 control) and the PSID comparison group (2490 control). We experiment on 10 original split subsets with 80% for training and 20% for testing [21,32].

- **Real datasets: Twins** The twins' dataset is derived from all twins born in the USA between 1989 and 1991. We set that treatment to $t = 1$ as the heavier twin and $t = 0$ as the lighter twin, and the outcome is the mortality of the children after one year. For each pair of twin records, we obtained 30 treatment variables related to parents, pregnancy, and birth [45]. We focus on twins weighing less than 2 kg without missing features, so the final dataset included 11 400 twin records. Because both the treatment and control outcomes can be observed, to simulate an observational study, we created selection biases as follows: $x \sim \text{Bern}\left(\text{sigmoid}\left(w^T X + n\right)\right)$, where $w^T \sim U\left((-0.1, 0.1)^{30\times 1}\right)$ and $\text{n} \sim N(0, 0.1)$ [45]. We experiment on the 10 original split subsets with 90% for training and 10% for testing [21,32,45].

- **Synthetic datasets** DR-CFR [31] and DeR-CFR [21] generate input samples based on independent normal distributions X_1, $X_2, \ldots, X_m \sim \mathcal{N}(0, 1)$. We generate 8 synthetic datasets with three variables m_Γ, m_Δ and m_Y, corresponding to $\Gamma(X)$, $\Delta(X)$, $Y(X)$, respectively, generated in $\{0, 8, 16\}$, and sample size $n = 3000$. For example, 16_16_16 denotes a dataset generated with 16 dimensions, individually representing the dimensions of instrumental, confounding, and risk factors. Thus, the total number of dimensions of X is $m = m_\Gamma + m_\Delta + m_Y + m_D$, where $m_D = 2$ represents the two noise variables. For treatment T, treatment sample t_i is generated from the $binomial(1, 1/(1+e^{-z}))$ distribution of the parameter $z = \frac{1}{10}\theta_t \times X_{\Gamma\Delta} + \varepsilon$. Then, generated binary outcomes of treatment and control groups are $Y^0 = \text{sign}\left(\max\left(0, z^0 - \bar{z}^0\right)\right)$ and $Y^1 = \text{sign}\left(\max\left(0, z^1 - \bar{z}^1\right)\right)$, respectively, where $z^0 = \frac{1}{10}\frac{\theta_{y0}\times X_{\Delta Y}}{m_\Delta + m_Y}$ and $z^1 = \frac{1}{10}\frac{\theta_{y1}\times X^2_{\Delta Y}}{m_\Delta + m_Y}$. Additionally, $\theta_t \sim \mathcal{U}\left((8, 16)^{m_\Gamma + m_\Delta}\right)$, $\theta_{y0}, \theta_{y1} \sim \mathcal{U}\left((8, 16)^{m_\Delta + m_Y}\right), \varepsilon \sim \mathcal{N}(0, 1)$ [21].

Evaluation Metrics. We mainly use indicators both Precision in Estimation of Heterogeneous Effect ($PEHE$) and absolute error in ATE (ϵ_{ATE}) to evaluate treatment effect, where $PEHE = \frac{1}{n}\sum_{i=1}^{n}([y_i'(1) - y_i'(0)] - [y_i(1) - y_i(0)])^2$ and $\epsilon_{ATE} = |ATE - \widehat{ATE}|$, $\widehat{ATE} = E[y_i'(1) - y_i'(0)]$.

Considering Jobs dataset cannot be evaluated by $PEHE$ and ϵ_{ATE} due to no ground truth [27], we use the policy risk (R_{pol}) and the absolute error in ATT (ϵ_{ATT}) to replace $PEHE$ and ϵ_{ATE}. The R_{pol} and ϵ_{ATT} are defined as follows [27]: $R_{\text{pol}} = 1 - (\mathbb{E}[y(1)|\pi(x) = 1]\,\mathcal{P}(\pi(x) = 1) + \mathbb{E}[y(0)|\pi(x) = 0]\,\mathcal{P}(\pi(x) = 0))$, where $\pi(x) = 1$ if $y(1) - y(0) > 0$ and $\pi(x) = 0$. $\epsilon_{ATT} = |ATT - \widehat{ATT}|$, $\widehat{ATT} = |T|^{-1}\sum_{i\in T}(y_i'(1) - y_i'(0))$.

Baselines. We compare our DRL_{ECB} with the following two groups of baselines. (1) Representation-based approach: CFR-MMD and CFR-WASS [27], CFR-ISW [28], SITE [29], DRRL [30]; (2) Decomposition-based approach: DR-CFR [31], TEDVAE [32], DeR-CFR [21], MIM-DRCFR [25], AutoIV [33]. The hyperparameter settings for the above baselines are shown in Table 3.

Implementation details. In our DRL_{ECB} model, the fully connected layer (FC) includes three independent representation layers with size = 200, ReLU activation function, and Stochastic Gradient Descent (SGD). This is because (1) FC can learn the global features. However, excessive increasing layers will result in the complexity of the model. Therefore, we balance the performance and time by setting the appropriate number of layers and neurons. (2) The ReLU function helps alleviate gradient vanishing and improve the model's stability and generalization. (3) SGD can only use a few samples to calculate the gradient and update parameters, making it very efficient in dealing with large-scale data.

6.2. Comparison on treatment effect estimation (RQ1)

Results on Real Datasets: We compare our DRL_{ECB} with 10 baselines, as shown in Table 4, on three real datasets. The baselines include the decomposition-based approaches TEDVAE [32], DR-CFR [31], DeR-CFR [21], MIM-DRCFR [25] and AutoIV [33] and representation-based approaches CFR-MMD and CFR-WASS [27], CFR-ISW [28], SITE [29] and DRRL [30]. The experiments are conducted under within-sample (train) and out-of-sample (test). Compared with the best result of baselines on within-sample and out-of-sample, (1) our DRL_{ECB} outperforms all baselines on IHDP dataset, with $PEHE$ decreasing 11% and 7%, and ϵ_{ATE} decreasing 37% and 1%. (2) On the Jobs dataset, our DRL_{ECB} outperforms all models. The R_{pol} declines 17% and 16% than the best DeR-CFR [21], and ϵ_{ATT} declines 52% and 27% than the best MIM-DRCFR [25]. (3) On the Twins dataset, our model reduces 83% and 62.5% on ϵ_{ATE}, respectively, while the performance on $PEHE$ is slightly worse. The above results are yielded for the following reasons: (1) In the IHDP and Jobs datasets in Table 4, DRL_{ECB} outperforms all baselines, indicating that the method of obtaining confounding representation through disentangling is effective. (2) DRL_{ECB} performs well in IHDP and Jobs, which indicates that $\Phi_\Delta(X)$ can effectively balance the quantities of treatment and control groups and eliminate the confounding bias. (3) Due to a large number of similar samples

Table 4
Experimental results of treatment effect estimation on Real datasets IHDP, Jobs and Twins.

Within-sample						
Datasets	IHDP (Mean ± Std)		Jobs (Mean ± Std)		Twins (Mean ± Std)	
Methods	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$R_{\text{pol}}(\downarrow)$	$\epsilon_{ATT}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
CFR-MMD	0.734 ± 0.042	0.285 ± 0.013	0.197 ± 0.005	0.043 ± 0.011	0.358 ± 0.023	0.028 ± 0.004
CFR-WASS	0.717 ± 0.132	0.292 ± 0.026	0.195 ± 0.013	0.041 ± 0.016	0.351 ± 0.017	0.022 ± 0.001
CFR-ISW	0.638 ± 0.071	0.215 ± 0.037	0.189 ± 0.058	0.041 ± 0.017	0.345 ± 0.024	0.037 ± 0.002
SITE	0.624 ± 0.062	0.264 ± 0.092	0.224 ± 0.005	0.067 ± 0.052	0.379 ± 0.021	0.033 ± 0.003
DRRL	0.632 ± 0.045	0.192 ± 0.052	0.221 ± 0.034	0.031 ± 0.034	0.324 ± 0.053	0.025 ± 0.004
DR-CFR	0.625 ± 0.066	0.247 ± 0.037	0.194 ± 0.016	0.057 ± 0.018	0.341 ± 0.013	0.017 ± 0.008
TEDVAE	0.596 ± 0.091	0.232 ± 0.060	0.199 ± 0.006	0.064 ± 0.026	0.311 ± 0.035	0.006 ± 0.002
DeR-CFR	0.507 ± 0.024	0.168 ± 0.028	0.187 ± 0.037	0.053 ± 0.084	**0.296 ± 0.011**	0.008 ± 0.002
MIM-DRCFR	0.518 ± 0.027	0.172 ± 0.046	0.190 ± 0.024	0.048 ± 0.022	0.327 ± 0.023	0.021 ± 0.007
AutoIV	0.637 ± 0.042	0.196 ± 0.034	0.196 ± 0.025	0.051 ± 0.026	0.331 ± 0.028	0.025 ± 0.009
DRL_{ECB}	**0.451 ± 0.042**	**0.105 ± 0.018**	**0.155 ± 0.029**	**0.023 ± 0.004**	0.312 ± 0.015	**0.001 ± 0.001**
w/t/l	10/0/0	10/0/0	10/0/0	10/0/0	8/0/2	10/0/0
Out-of-sample						
Datasets	IHDP (Mean ± Std)		Jobs (Mean ± Std)		Twins (Mean ± Std)	
Methods	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$R_{\text{pol}}(\downarrow)$	$\epsilon_{ATT}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
CFR-MMD	0.782 ± 0.042	0.312 ± 0.033	0.217 ± 0.061	0.079 ± 0.034	0.363 ± 0.041	0.015 ± 0.003
CFR-WASS	0.769 ± 0.063	0.324 ± 0.025	0.221 ± 0.037	0.093 ± 0.015	0.357 ± 0.087	0.022 ± 0.005
CFR-ISW	0.702 ± 0.036	0.231 ± 0.083	0.236 ± 0.018	0.069 ± 0.023	0.348 ± 0.092	0.039 ± 0.004
SITE	0.681 ± 0.014	0.347 ± 0.113	0.237 ± 0.026	0.072 ± 0.031	0.384 ± 0.056	0.040 ± 0.006
DRRL	0.685 ± 0.028	0.217 ± 0.036	0.257 ± 0.024	0.047 ± 0.022	0.367 ± 0.036	0.032 ± 0.004
DR-CFR	0.684 ± 0.015	0.266 ± 0.086	0.205 ± 0.048	0.081 ± 0.025	0.350 ± 0.018	0.014 ± 0.007
TEDVAE	0.617 ± 0.026	0.258 ± 0.049	0.218 ± 0.027	0.072 ± 0.042	0.318 ± 0.006	0.008 ± 0.002
DeR-CFR	0.602 ± 0.068	0.185 ± 0.023	0.192 ± 0.026	0.066 ± 0.037	**0.303 ± 0.068**	0.009 ± 0.002
MIM-DRCFR	0.594 ± 0.039	0.195 ± 0.051	0.220 ± 0.031	0.060 ± 0.033	0.347 ± 0.045	0.029 ± 0.009
AutoIV	0.692 ± 0.059	0.224 ± 0.035	0.214 ± 0.022	0.067 ± 0.038	0.355 ± 0.032	0.029 ± 0.011
DRL_{ECB}	**0.552 ± 0.039**	**0.183 ± 0.025**	**0.162 ± 0.007**	**0.044 ± 0.007**	0.324 ± 0.031	**0.003 ± 0.004**
w/t/l	10/0/0	10/0/0	10/0/0	10/0/0	8/0/2	10/0/0

The win/tie/loss counts for DRL_{ECB} are summarized in the last row, abbreviated as w/t/l.

Table 5
Experimental results of $PEHE$ estimation on Synthetic datasets with different dimensions of $\Gamma(X)$, $\Delta(X)$, $Y(X)$ (m_Γ, m_Δ, m_Y).

Datasets	DR-CFR (Mean ± Std)	TEDVAE (Mean ± Std)	DeR-CFR (Mean ± Std)	MIM-DRCFR (Mean ± Std)	AutoIV (Mean ± Std)	DRL_{ECB} (Mean ± Std)
Within-sample	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$
8_8_16	0.581 ± 0.027	0.545 ± 0.078	0.416 ± 0.027	0.534 ± 0.031	0.567 ± 0.044	**0.392 ± 0.019**
16_8_8	0.548 ± 0.039	0.511 ± 0.037	0.425 ± 0.047	0.506 ± 0.033	0.541 ± 0.038	**0.403 ± 0.024**
8_16_8	0.479 ± 0.058	0.431 ± 0.029	0.375 ± 0.032	0.441 ± 0.029	0.468 ± 0.035	**0.335 ± 0.019**
16_0_16	0.515 ± 0.065	0.458 ± 0.034	0.423 ± 0.033	0.460 ± 0.028	0.503 ± 0.037	**0.398 ± 0.017**
8_16_16	0.582 ± 0.044	0.527 ± 0.039	0.481 ± 0.027	0.496 ± 0.030	0.514 ± 0.034	**0.435 ± 0.034**
16_8_16	0.527 ± 0.057	0.574 ± 0.028	0.479 ± 0.014	0.501 ± 0.031	0.547 ± 0.041	**0.446 ± 0.028**
16_16_8	0.553 ± 0.036	0.518 ± 0.037	0.454 ± 0.029	0.511 ± 0.029	0.532 ± 0.035	**0.449 ± 0.034**
16_16_16	0.534 ± 0.024	0.509 ± 0.035	**0.417 ± 0.015**	0.518 ± 0.032	0.531 ± 0.037	0.437 ± 0.018
Avg.	0.540 ± 0.044	0.509 ± 0.039	0.435 ± 0.027	0.496 ± 0.030	0.525 ± 0.038	**0.411 ± 0.024**
w/t/l	–	–	–	–	–	7/0/1
Out-of-sample	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$	$PEHE(\downarrow)$
8_8_16	0.602 ± 0.127	0.585 ± 0.045	**0.441 ± 0.057**	0.586 ± 0.044	0.614 ± 0.051	0.454 ± 0.029
16_8_8	0.624 ± 0.108	0.601 ± 0.086	0.514 ± 0.071	0.562 ± 0.035	0.602 ± 0.039	**0.470 ± 0.031**
8_16_8	0.528 ± 0.078	0.513 ± 0.036	0.457 ± 0.041	0.492 ± 0.038	0.516 ± 0.041	**0.405 ± 0.037**
16_0_16	0.553 ± 0.025	0.498 ± 0.057	0.471 ± 0.023	0.511 ± 0.033	0.578 ± 0.043	**0.453 ± 0.036**
8_16_16	0.609 ± 0.035	0.584 ± 0.045	0.551 ± 0.035	0.519 ± 0.035	0.541 ± 0.038	**0.502 ± 0.045**
16_8_16	0.587 ± 0.072	0.614 ± 0.067	0.483 ± 0.027	0.548 ± 0.029	0.579 ± 0.036	**0.478 ± 0.024**
16_16_8	0.601 ± 0.058	0.553 ± 0.048	0.514 ± 0.031	0.558 ± 0.030	0.549 ± 0.031	**0.471 ± 0.042**
16_16_16	0.579 ± 0.048	0.558 ± 0.055	**0.495 ± 0.024**	0.557 ± 0.036	0.582 ± 0.042	**0.495 ± 0.022**
Avg.	0.585 ± 0.039	0.563 ± 0.054	0.502 ± 0.038	0.542 ± 0.035	0.570 ± 0.040	**0.466 ± 0.033**
w/t/l	–	–	–	–	–	6/1/1

in Twins dataset, our DRL_{ECB} cannot present significantly outstanding.

Results on Synthetic Datasets: To analyze the performance on synthetic datasets, we compare our DRL_{ECB} with five competitive baselines, namely DR-CFR [31], TEDVAE [32], DeR-CFR [21], MIM-DRCFR [25] and AutoIV [33]. From Tables 5 and 6, we can observe that our model rises 6% and 43% on $PEHE$ and ϵ_{ATE}, respectively, on within-sample, and rises 7% and 36% on the out-of-sample relative to the average of the best baseline. The main reasons why DRL_{ECB} outperforms its rivals include: (1) The benefit of separating confounders and eliminating bias by controlling for confounders are evident when both instrumental and risk factors are present in the data. (2) DeR-CFR [21] mainly adopts the MMD method to realize the decomposition; however, the MMD has high computational complexity and is easy to overfit when facing high-dimensional.

Table 6
Experimental results of ϵ_{ATE} on Synthetic datasets with different dimensions of $\Gamma(X)$, $\Delta(X)$, $Y(X)$ (m_Γ, m_Δ, m_Y).

Datasets	DR-CFR (Mean ± Std)	TEDVAE (Mean ± Std)	DeR-CFR (Mean ± Std)	MIM-DRCFR (Mean ± Std)	AutoIV (Mean ± Std)	DRL_{ECB} (Mean ± Std)
Within-sample	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
8_8_16	0.034 ± 0.007	0.040 ± 0.011	0.024 ± 0.006	0.029 ± 0.009	0.035 ± 0.011	**0.017 ± 0.005**
16_8_8	0.037 ± 0.005	0.025 ± 0.008	0.023 ± 0.005	0.030 ± 0.007	0.037 ± 0.013	**0.009 ± 0.003**
8_16_8	0.048 ± 0.013	0.034 ± 0.010	0.025 ± 0.007	0.028 ± 0.010	0.033 ± 0.010	**0.011 ± 0.004**
16_0_16	0.046 ± 0.011	0.048 ± 0.016	0.034 ± 0.008	0.035 ± 0.013	0.040 ± 0.012	**0.027 ± 0.008**
8_16_16	0.021 ± 0.006	0.028 ± 0.007	0.015 ± 0.004	0.019 ± 0.008	0.025 ± 0.009	**0.002 ± 0.001**
16_8_16	0.043 ± 0.017	0.043 ± 0.015	0.032 ± 0.005	0.035 ± 0.007	0.039 ± 0.009	**0.021 ± 0.008**
16_16_8	0.067 ± 0.009	0.064 ± 0.013	0.057 ± 0.009	0.060 ± 0.012	0.061 ± 0.014	**0.041 ± 0.010**
16_16_16	0.025 ± 0.015	0.028 ± 0.009	0.014 ± 0.004	0.019 ± 0.008	0.025 ± 0.010	**0.006 ± 0.002**
Avg.	0.040 ± 0.010	0.035 ± 0.011	0.028 ± 0.006	0.032 ± 0.009	0.037 ± 0.011	**0.016 ± 0.005**
w/t/l	–	–	–	–	–	8/0/0
Out-of-sample	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
8_8_16	0.037 ± 0.008	0.042 ± 0.016	0.026 ± 0.008	0.032 ± 0.011	0.040 ± 0.014	**0.014 ± 0.008**
16_8_8	0.044 ± 0.007	0.031 ± 0.007	0.031 ± 0.005	0.035 ± 0.009	0.041 ± 0.013	**0.017 ± 0.009**
8_16_8	0.058 ± 0.005	0.045 ± 0.008	0.032 ± 0.006	0.033 ± 0.010	0.039 ± 0.011	**0.026 ± 0.007**
16_0_16	0.051 ± 0.011	0.049 ± 0.011	0.028 ± 0.008	0.040 ± 0.008	0.045 ± 0.012	**0.016 ± 0.007**
8_16_16	0.032 ± 0.009	0.031 ± 0.009	0.019 ± 0.009	0.024 ± 0.010	0.030 ± 0.011	**0.008 ± 0.002**
16_8_16	0.051 ± 0.017	0.050 ± 0.012	0.034 ± 0.005	0.040 ± 0.011	0.030 ± 0.012	**0.028 ± 0.008**
16_16_8	0.053 ± 0.015	0.044 ± 0.015	0.041 ± 0.007	0.045 ± 0.012	0.048 ± 0.013	**0.021 ± 0.007**
16_16_16	0.038 ± 0.012	0.048 ± 0.017	0.029 ± 0.011	0.030 ± 0.008	0.033 ± 0.010	**0.025 ± 0.009**
Avg.	0.068 ± 0.014	0.042 ± 0.011	0.030 ± 0.007	0.035 ± 0.010	0.038 ± 0.012	**0.019 ± 0.007**
w/t/l	–	–	–	–	–	8/0/0

Table 7
Ablation experiments (Mean ± Std) on effectiveness of disentangling X into $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$ on Real datasets.

Dataset		DRL_{ECB}	Without disentangling X	Without $\Phi_\Gamma(X)$	Without $\Phi_\Delta(X)$	Without $\Phi_Y(X)$
IHDP ($PEHE$)	Within-sample	**0.451 ± 0.042**	0.574 ± 0.036	0.521 ± 0.044	0.517 ± 0.043	0.518 ± 0.043
	Out-of-sample	**0.552 ± 0.039**	0.652 ± 0.047	0.611 ± 0.055	0.611 ± 0.048	0.611 ± 0.043
Jobs (R_{pol})	Within-sample	**0.155 ± 0.029**	0.185 ± 0.027	0.168 ± 0.014	0.164 ± 0.025	0.170 ± 0.008
	Out-of-sample	**0.162 ± 0.007**	0.195 ± 0.017	0.169 ± 0.012	0.172 ± 0.008	0.174 ± 0.024
Twins ($PEHE$)	Within-sample	**0.312 ± 0.015**	0.337 ± 0.005	0.323 ± 0.068	0.326 ± 0.045	0.332 ± 0.054
	Out-of-sample	**0.324 ± 0.031**	0.348 ± 0.013	0.334 ± 0.038	0.341 ± 0.029	0.338 ± 0.031

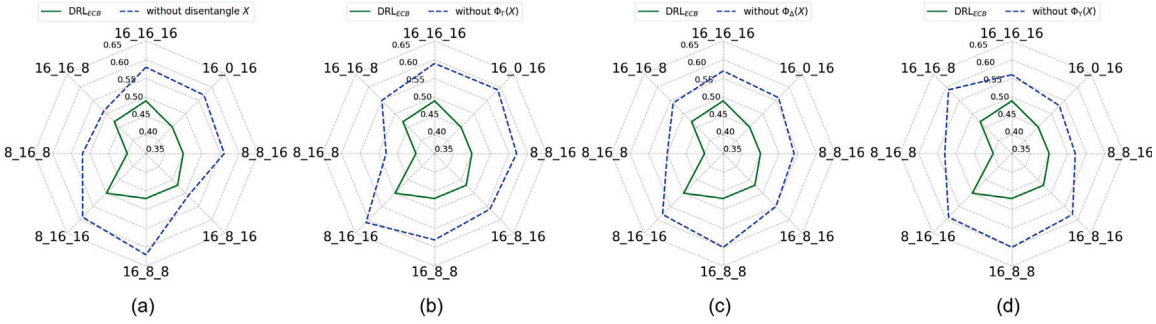


Fig. 4. Ablation experiments (Mean ± Std) on effectiveness of disentangling X into $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$ on Synthetic datasets. (a) without disentangled X, (b) without $\Phi_\Gamma(X)$, (c) without $\Phi_\Delta(X)$, and (d) without $\Phi_Y(X)$ on 8 synthetic datasets, respectively. The vertices in graphs are identified by (m_Γ_m_Δ_m_Y), which denote the dimensions of synthetic dataset on $\{\Gamma(X), \Delta(X), Y(X)\}$, respectively. The polygon's radius ranges from 0.35 to 0.65. For $PEHE$, the smaller the value (near the center), the better the performance of the model.

Table 8
Ablation experiments (Mean ± Std) with regularization and its variants on Real and Synthetic datasets.

Categories	Dataset	DRL_{ECB}	Without regularizing	Without minimizing MI $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$	Without minimizing MI $\Phi_\Gamma(X)$ and $\Phi_Y(X)$	Without minimizing MI $\Phi_\Delta(X)$ and $\Phi_Y(X)$
Real	IHDP ($PEHE$)	**0.552 ± 0.039**	0.624 ± 0.024	0.605 ± 0.018	0.601 ± 0.020	0.594 ± 0.017
	Jobs (R_{pol})	**0.162 ± 0.007**	0.175 ± 0.011	0.169 ± 0.008	0.168 ± 0.007	0.168 ± 0.005
	Twins ($PEHE$)	**0.324 ± 0.031**	0.336 ± 0.028	0.331 ± 0.022	0.330 ± 0.019	0.329 ± 0.018
Synthetic	8_16_16 ($PEHE$)	**0.502 ± 0.045**	0.668 ± 0.035	0.618 ± 0.027	0.607 ± 0.029	0.592 ± 0.031
	16_8_16 ($PEHE$)	**0.478 ± 0.024**	0.584 ± 0.026	0.514 ± 0.028	0.521 ± 0.029	0.505 ± 0.022
	16_16_8 ($PEHE$)	**0.471 ± 0.042**	0.583 ± 0.038	0.512 ± 0.029	0.529 ± 0.032	0.504 ± 0.033
	16_16_16 ($PEHE$)	**0.495 ± 0.022**	0.601 ± 0.025	0.548 ± 0.017	0.550 ± 0.023	0.547 ± 0.021

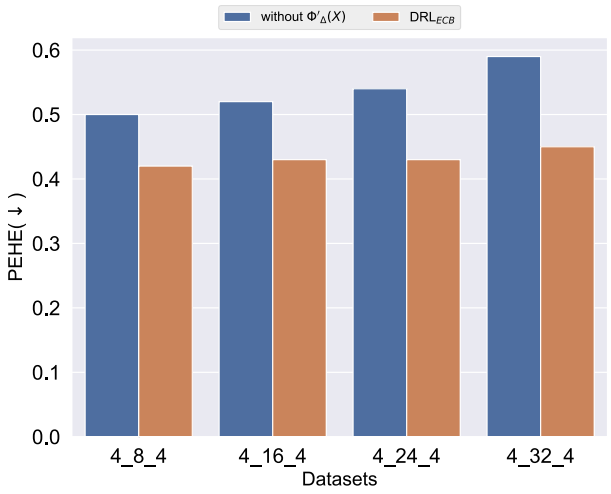


Fig. 5. The impact of $\Phi'_\Delta(X)$ on datasets with different dimensional $\Delta(X)$ (m_Δ).

Table 9
Results (Mean ± Std) of without $\Phi'_\Delta(X)$ on IHDP, Jobs, Twins and Synthetic datasets.

Categories	Dataset	DRL_{ECB}	Without $\Phi'_\Delta(X)$
Real	IHDP ($PEHE$)	**0.552 ± 0.039**	0.607 ± 0.024
	Jobs (R_{pol})	**0.162 ± 0.007**	0.172 ± 0.016
	Twins ($PEHE$)	**0.324 ± 0.031**	0.331 ± 0.024
Synthetic	8_16_16 ($PEHE$)	**0.502 ± 0.045**	0.564 ± 0.037
	16_8_16 ($PEHE$)	**0.478 ± 0.034**	0.517 ± 0.026
	16_16_8 ($PEHE$)	**0.471 ± 0.039**	0.513 ± 0.038

6.3. Ablation study

This section shows the effects on disentanglement X, $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$ and $\Phi_Y(X)$ and regularization on real and synthetic datasets, respectively.

6.3.1. Effectiveness analysis on disentangling X into $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$ and $\Phi_Y(X)$ (RQ2-1)

From Table 7 and Fig. 4, we can draw the following results: (1) In Table 7, we can see that it performs the worst on the real datasets without disentangling X. Fig. 4(a) shows that the effect without disentanglement X is significantly lower than that of DRL_{ECB}, which denotes that learning disentangled representations is necessary and effective. (2) In Table 7 and Fig. 4(b)–(d), we remove $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$ in DRL_{ECB} and compare them with DRL_{ECB}, respectively. We can note that the effect of removing any of these representations is lower than the original DRL_{ECB}, so $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$ and $\Phi_Y(X)$ in DRL_{ECB} are crucial, which affects the performance of our treatment effect estimation.

6.3.2. Effectiveness analysis on regularizing $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, and $\Phi_Y(X)$ (RQ2-2)

To verify the importance of additional independence enhancement by regularizing $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$, we vary the regularization from four variants. i.e., without regularizing, without minimizing MI $\Phi_\Gamma(X)$ and $\Phi_\Delta(X)$, without minimizing MI $\Phi_\Gamma(X)$ and $\Phi_Y(X)$, without minimizing MI $\Phi_\Delta(X)$ and $\Phi_Y(X)$. In Table 8, we remove all regularizations in the DRL_{ECB} and the regularizations between the three representations on the real and synthetic datasets, respectively. Comparing the DRL_{ECB} with the model that removes regularization separately, it can be seen that removing any regularization between representations as well as removing all regularization is less effective than the DRL_{ECB}. Minimizing MI improves the independence between the three representations, thereby enhancing the performance of treatment effect estimation.

6.4. Effectiveness analysis on $\Phi'_\Delta(X)$ (RQ3)

This section presents the effectiveness of confounding balance $\Phi'_\Delta(X)$ on real and synthetic datasets. From Table 9, when $\Phi'_\Delta(X)$ is removed from DRL_{ECB}, the results turn worse than original DRL_{ECB}. It denotes $\Phi'_\Delta(X)$ is an important part of estimating causal effects. Fig. 5 shows the results of $\Phi'_\Delta(X)$ on synthetic datasets with $\Delta(X)$ of different dimensions (m_Δ). We can observe that when ignoring $\Phi'_\Delta(X)$, as m_Δ increases, the confounding bias also increases, which leads to decreased performance. Conversely, when $\Phi'_\Delta(X)$ is fitted, the performance is significantly improved on different m_Δ. It indicates that $\Phi'_\Delta(X)$ can effectively deal with the confounding bias caused by $\Delta(X)$, and improve the performance of the model.

6.5. Effect of hyperparameter β

In our experiments, we set hyperparameter β to 1.0. To estimate the impact of hyperparameter β on model performance, we set the typical value of β from $\{0.1, 0.5, 1.0, 1.5, 3.0, 4.0, 5.0\}$ on different datasets. The experimental results are shown in Table 10. We can observe that although the performance of the DRL_{ECB} model is slightly better when β is 1.0, it varies less on different β and a smaller MAD ratio indicates less data variation. This further indicates that β does not cause significant fluctuations in the performance of DRL_{ECB}. This is mainly because the balancing term $\mathcal{L}_\Delta^{bal}$ is optimized based on the input variables X and T without relying on the outcome Y, which allows model DRL_{ECB} to automatically adjust the balancing term $\mathcal{L}_\Delta^{bal}$. Even under different values of β, DRL_{ECB} can still learn the balancing representation stably, thus ensuring the accuracy of downstream estimations.

6.6. Training time analysis

To analyze the complexity and training time of the methods, we repeated the above baseline and DRL_{ECB} methods 10 times on real and synthetic datasets, respectively, and compared the average training time for a single execution of the baseline and DRL_{ECB} methods, respectively, as shown in Table 11.

From Table 11, we can get the following results: (1) In terms of execution time, the representation-based methods is overall smaller than the decomposition-based methods, because the implementation of the disentanglement functions can effectively improve the accuracy of the effect estimation, but the same implementation of the disentanglement will increase the complexity of the methods and the cost of training, so it leads to the increase of the execution time of the decomposition-based methods; (2) Among the decomposition-based methods, the single execution time of DRL_{ECB}, which is obviously smaller than other decomposition-based methods, and can be comparable to the representation-based methods, because the disentanglement by MI in DRL_{ECB}, for each positive sample pair(i.e. $(\phi_\Gamma^c(\boldsymbol{x}_i), \boldsymbol{t}_i)$) replaces the computation the mean of the probabilities of all negative sample pairs(i.e $\frac{1}{N}\sum_{j=1}^{N}\log q_{\theta_{\Gamma T}}(\boldsymbol{t}_j|\phi_\Gamma^c(\boldsymbol{x}_i))$) with the log probability $\log q_{\theta_{\Gamma T}}(\boldsymbol{t}_{k_i}|\phi_\Gamma^c(\boldsymbol{x}_i))$ of random sampling of a negative sample pair(i.e. $(\phi_\Gamma^c(\boldsymbol{x}_i), \boldsymbol{t}_{k_i})$, $k_i \in N$, randomly sampling k_i from N), which makes it possible to improve the effect estimation accuracy while shortening the execution time of the method and reducing the training cost of the model.

6.7. A case study

In this paper, we utilize real-world social media dataset BlogCatalog[1] [46] to estimate treatment effects by DRL_{ECB}. BlogCatalog dataset

[1] https://github.com/rguo12/network-deconfounder-wsdm20/tree/master/datasets/BlogCatalog1.

Table 10
Experimental results of DRL_{ECB} with different β on Real and Synthetic datasets. MAD is the mean absolute deviation of Means.

β	IHDP (Mean ± Std)		Jobs (Mean ± Std)		Twins (Mean ± Std)		16_16_16 (Mean ± Std)	
Within-sample	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$R_{pol}(\downarrow)$	$\epsilon_{ATT}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
0.1	0.462 ± 0.044	0.111 ± 0.012	0.162 ± 0.033	0.030 ± 0.007	0.357 ± 0.021	0.003 ± 0.001	0.452 ± 0.020	0.013 ± 0.006
0.5	0.458 ± 0.032	0.109 ± 0.019	0.160 ± 0.030	0.026 ± 0.003	0.330 ± 0.018	0.002 ± 0.001	**0.436 ± 0.026**	0.015 ± 0.005
1.0	**0.451 ± 0.042**	0.105 ± 0.018	0.155 ± 0.029	**0.023 ± 0.004**	**0.312 ± 0.015**	**0.001 ± 0.001**	0.437 ± 0.018	**0.006 ± 0.002**
1.5	0.452 ± 0.035	**0.103 ± 0.015**	**0.154 ± 0.032**	0.024 ± 0.004	0.318 ± 0.014	**0.001 ± 0.001**	0.439 ± 0.023	0.011 ± 0.004
3.0	0.467 ± 0.040	0.122 ± 0.016	0.159 ± 0.031	0.032 ± 0.006	0.345 ± 0.022	0.002 ± 0.001	0.448 ± 0.022	0.013 ± 0.003
4.0	0.479 ± 0.038	0.135 ± 0.017	0.164 ± 0.034	0.037 ± 0.007	0.375 ± 0.019	0.003 ± 0.002	0.462 ± 0.025	0.014 ± 0.005
5.0	0.485 ± 0.045	0.152 ± 0.023	0.172 ± 0.035	0.038 ± 0.009	0.394 ± 0.020	0.004 ± 0.002	0.471 ± 0.029	0.016 ± 0.007
MAD	0.010	0.069	0.004	0.005	0.024	0.001	0.011	0.002
Out-of-sample	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$R_{pol}(\downarrow)$	$\epsilon_{ATT}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
0.1	0.589 ± 0.047	0.201 ± 0.029	0.175 ± 0.010	0.061 ± 0.009	0.369 ± 0.033	0.005 ± 0.004	0.503 ± 0.023	0.029 ± 0.007
0.5	0.556 ± 0.040	0.187 ± 0.030	0.168 ± 0.008	0.058 ± 0.003	0.326 ± 0.018	0.004 ± 0.003	**0.482 ± 0.026**	0.027 ± 0.005
1.0	**0.552 ± 0.039**	**0.183 ± 0.025**	**0.162 ± 0.007**	**0.044 ± 0.007**	**0.324 ± 0.031**	**0.003 ± 0.004**	0.495 ± 0.022	**0.025 ± 0.009**
1.5	0.559 ± 0.044	0.190 ± 0.034	0.168 ± 0.009	0.055 ± 0.006	0.327 ± 0.024	0.004 ± 0.003	0.488 ± 0.024	0.026 ± 0.004
3.0	0.571 ± 0.042	0.198 ± 0.031	0.175 ± 0.010	0.059 ± 0.008	0.369 ± 0.025	0.004 ± 0.003	0.496 ± 0.023	0.028 ± 0.006
4.0	0.583 ± 0.043	0.201 ± 0.032	0.181 ± 0.013	0.061 ± 0.010	0.387 ± 0.030	0.005 ± 0.003	0.504 ± 0.023	0.029 ± 0.007
5.0	0.602 ± 0.046	0.212 ± 0.033	0.189 ± 0.015	0.064 ± 0.011	0.412 ± 0.032	0.005 ± 0.004	0.512 ± 0.025	0.030 ± 0.008
MAD	0.016	0.008	0.007	0.005	0.029	0.001	0.008	0.001

Table 11
Average training time(s) in a single execution of DRL_{ECB} and baselines on Real and Synthetic datasets.

Methods	IHDP	Jobs	Twins	8_8_8	8_16_16	16_8_16	16_16_8	16_16_16
CFR-MMD	36.7 ± 2.4	40.5 ± 3.1	67.2 ± 4.5	42.4 ± 2.3	43.1 ± 1.9	43.9 ± 2.0	42.6 ± 2.4	45.7 ± 3.6
CFR-WASS	40.8 ± 3.1	63.5 ± 3.6	110.6 ± 4.2	53.3 ± 2.8	54.5 ± 3.2	54.1 ± 2.8	54.8 ± 2.9	55.2 ± 3.9
CFR-ISW	50.1 ± 2.6	52.4 ± 2.9	88.0 ± 4.2	52.7 ± 2.4	53.3 ± 2.5	54.2 ± 2.3	53.8 ± 1.8	54.5 ± 2.8
SITE	55.2 ± 2.9	52.3 ± 3.1	85.0 ± 3.5	73.2 ± 3.1	74.5 ± 2.8	73.8 ± 2.7	73.4 ± 3.0	74.6 ± 3.2
DRRL	60.1 ± 2.7	62.4 ± 2.4	95.4 ± 3.2	69.7 ± 2.7	70.5 ± 2.8	69.9 ± 2.3	71.3 ± 3.2	72.8 ± 3.4
DR-CFR	75.3 ± 2.3	142.7 ± 3.7	165.4 ± 4.8	74.2 ± 3.1	75.8 ± 2.5	75.1 ± 3.2	74.7 ± 3.4	76.7 ± 2.8
TEDVAE	77.4 ± 2.1	100.6 ± 3.3	178.3 ± 4.6	95.4 ± 3.2	96.2 ± 3.5	95.8 ± 3.8	96.1 ± 3.5	97.5 ± 3.7
DeR-CFR	76.1 ± 2.4	150.4 ± 4.7	230.2 ± 5.5	106.8 ± 3.3	108.5 ± 3.5	109.1 ± 3.2	108.7 ± 3.0	110.5 ± 3.6
MIM-DRCFR	79.4 ± 2.8	138.5 ± 3.6	195.4 ± 5.2	85.4 ± 3.1	86.7 ± 3.3	85.8 ± 3.1	86.4 ± 3.2	87.1 ± 3.4
AutoIV	82.4 ± 3.2	96.5 ± 3.7	175.1 ± 4.1	82.3 ± 0.0	83.4 ± 2.8	82.7 ± 3.0	83.1 ± 3.4	85.5 ± 2.9
DRL_{ECB}	64.7 ± 2.1	52.3 ± 1.8	140.2 ± 2.7	60.1 ± 1.9	56.7 ± 2.0	53.7 ± 2.1	54.5 ± 2.3	58.6 ± 2.2

includes 5196 bloggers. The treatment variables (X) are a description based on the topic of the blogger's posted blog with a total of 8189 features. The treatment (T) indicates whether the blogs are browsed on mobile devices or desktops. To observe the relationships between browsing preferences and reading devices, we set a treatment group ($T = 1$) when using mobile devices and a control group ($T = 0$) when using desktops. The outcome (Y) is the readers' perception of the blogger. We aim to analyze the treatment effect of whether getting more views on mobile devices/desktops has on bloggers' perceptions. We randomly select instances 80% for training and 20% for testing, respectively.

Table 12 shows the evaluation of the treatment effect of DRL_{ECB} and 10 baselines on BlogCatalog dataset. We can see that our DRL_{ECB} improves 13% and 23% on within-sample and 5% and 51% on out-of-sample compared to its baselines on $PEHE$ and ϵ_{ATE} metrics respectively. From Table 12, we can conclude that DRL_{ECB} still shows better performance and efficiency with a high dimensionality of X compared to the baseline, demonstrating the generalizability of our approach.

7. Conclusion

This paper explores a new disentangled representation model, DRL_{ECB}, with excluding confounding bias to estimate treatment effects in observational data. First, we disentangle the treatment variables X into three representations $\Phi_\Gamma(X)$, $\Phi_\Delta(X)$, $\Phi_Y(X)$ by calculating the correlation of three factors $\Gamma(X), \Delta(X), Y(X)$ with the treatment T and the outcome Y by MI; Second, we balance confounding representation $\Phi_\Delta(X)$ between the treatment and control groups by minimizing the squared deviation between the reweighted treatment and control groups to eliminate the confounding bias caused by confounders in observational data; Finally, we constructs regression network models h^0 and h^1 based on the treatment and control groups, respectively, to estimate treatment effects. Experimental results show that DRL_{ECB} performs better than the state-of-the-art method in most cases.

Although our DRL_{ECB} can work well in estimating causal effects in structured data with confounding bias, estimating causal effects in unstructured data is still a challenging and open problem. In the future, we plan to systematically study the mechanisms of causal effect estimation and exclude confounding bias under open set unstructured, multi-sources, and multimodal data.

CRediT authorship contribution statement

Dianlong You: Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Dongyan Wang:** Writing – review & editing, Writing – original draft, Methodology, Formal analysis, Data curation, Conceptualization. **Bingxin Liu:** Resources, Investigation. **Xiaoyi Ge:** Resources, Investigation. **Di Wu:** Resources, Funding acquisition. **Xindong Wu:** Supervision, Methodology.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work has been partially supported by grants from the National Natural Science Foundation of China No. 62276226, No. 62176070, and 62120106008; S&T Program of Hebei, No. 236Z7725G; Project of Hebei Key Laboratory of Software Engineering No. 22567637H.

Table 12
Results of treatment effect estimation on BlogCatalog.

Within-sample		
Datasets	BlogCatalog (Mean ± Std)	
Methods	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
CFR-MMD	11.547 ± 5.124	4.358 ± 2.374
CFR-WASS	10.856 ± 5.785	4.139 ± 2.341
CFR-ISW	8.945 ± 3.785	4.639 ± 1.941
SITE	7.542 ± 3.146	2.356 ± 1.292
DRRL	5.362 ± 2.348	1.953 ± 1.044
DR-CFR	7.962 ± 3.125	2.963 ± 1.945
TEDVAE	6.174 ± 2.486	1.575 ± 0.978
DeR-CFR	8.423 ± 5.756	2.846 ± 1.683
MIM-DRCFR	6.652 ± 1.821	1.354 ± 0.869
AutoIV	8.674 ± 2.412	2.741 ± 1.057
DRL_{ECB}	**4.673 ± 3.690**	**1.041 ± 0.845**
w/t/l	10/0/0	10/0/0
Out-of-sample		
Datasets	BlogCatalog (Mean ± Std)	
Methods	$PEHE(\downarrow)$	$\epsilon_{ATE}(\downarrow)$
CFR-MMD	12.738 ± 6.454	5.471 ± 2.121
CFR-WASS	11.736 ± 5.012	5.324 ± 2.251
CFR-ISW	10.645 ± 5.457	5.349 ± 2.218
SITE	8.629 ± 2.894	3.184 ± 1.255
DRRL	7.042 ± 2.548	2.840 ± 1.088
DR-CFR	8.841 ± 5.426	3.572 ± 2.701
TEDVAE	7.531 ± 3.454	2.261 ± 1.232
DeR-CFR	9.145 ± 3.754	3.572 ± 1.664
MIM-DRCFR	7.642 ± 2.024	2.143 ± 1.021
AutoIV	9.275 ± 3.041	3.314 ± 1.866
DRL_{ECB}	**6.667 ± 4.215**	**1.053 ± 0.876**
w/t/l	10/0/0	10/0/0

Data availability

Data will be made available on request.

References

[1] L. Li, P. Shi, Q. Fan, W. Zhong, Causal effect estimation with censored outcome and covariate selection, Statist. Probab. Lett. 204 (2024) 109933.
[2] M.R. Heydari, S. Salehkaleybar, K. Zhang, Adversarial orthogonal regression: Two non-linear regressions for causal inference, Neural Netw. 143 (2021) 66–73.
[3] R. Cai, W. Chen, Z. Yang, S. Wan, C. Zheng, X. Yang, J. Guo, Long-term causal effects estimation via latent surrogates representation learning, Neural Netw. 176 (2024) 106336.
[4] L. Jia, T.W. Chow, Y. Yuan, Causal disentanglement domain generalization for time-series signal fault diagnosis, Neural Netw. 172 (2024) 106099.
[5] E. Igelström, P. Craig, J. Lewsey, J. Lynch, A. Pearce, S.V. Katikireddi, Causal inference and effect estimation using observational data, J. Epidemiol. Community Health 76 (11) (2022) 960–966.
[6] P. Jiao, H. Chen, H. Tang, Q. Bao, L. Zhang, Z. Zhao, H. Wu, Contrastive representation learning on dynamic networks, Neural Netw. 174 (2024) 106240.
[7] S. Gupta, Z. Lipton, D. Childers, Efficient online estimation of causal effects by deciding what to observe, Adv. Neural Inf. Process. Syst. 34 (2021) 20995–21007.
[8] N. Kallus, M. Uehara, Double reinforcement learning for efficient off-policy evaluation in markov decision processes, J. Mach. Learn. Res. 21 (167) (2020) 1–63.
[9] Y. Tian, K. Bai, X. Yu, S. Zhu, Causal multi-label learning for image classification, Neural Netw. 167 (2023) 626–637.
[10] D. Cheng, J. Li, L. Liu, J. Liu, T.D. Le, Data-driven causal effect estimation based on graphical causal modelling: A survey, ACM Comput. Surv. 56 (5) (2024) 1–37.
[11] P. Cui, S. Athey, Stable learning establishes some common ground between causal inference and machine learning, Nat. Mach. Intell. 4 (2) (2022) 110–115.
[12] S. Zhang, X. Feng, W. Fan, W. Fang, F. Feng, W. Ji, S. Li, L. Wang, S. Zhao, Z. Zhao, et al., Video-audio domain generalization via confounder disentanglement, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 15322–15330.
[13] P. Manomaisaowapak, J. Songsiri, Joint learning of multiple granger causal networks via non-convex regularizations: Inference of group-level brain connectivity, Neural Netw. 149 (2022) 157–171.
[14] C. Li, Y. Mao, S. Liang, J. Li, Y. Wang, Y. Guo, Deep causal learning for pancreatic cancer segmentation in CT sequences, Neural Netw. 175 (2024) 106294.
[15] Y. Yang, B. Xu, S. Shen, F. Shen, J. Zhao, Operation-aware neural networks for user response prediction, Neural Netw. 121 (2020) 161–168.
[16] H. Su, Z. Du, J. Li, L. Zhu, K. Lu, Cross-domain adaptative learning for online advertisement customer lifetime value prediction, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 4605–4613.
[17] G.J. Hitsch, S. Misra, W.W. Zhang, Heterogeneous treatment effects and optimal targeting policy evaluation, Quant. Mark. Econ. 22 (2) (2024) 115–168.
[18] H. Zhou, S. Li, G. Jiang, J. Zheng, D. Wang, Direct heterogeneous causal learning for resource allocation problems in marketing, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 37, 2023, pp. 5446–5454.
[19] T. Tezuka, M. Kuroki, An unbiased estimator of the causal effect on the variance based on the back-door criterion in Gaussian linear structural equation models, J. Multivariate Anal. 197 (2023) 105201.
[20] R. Shanmugam, Elements of causal inference: foundations and learning algorithms, 2018.
[21] A. Wu, J. Yuan, K. Kuang, B. Li, R. Wu, Q. Zhu, Y. Zhuang, F. Wu, Learning decomposed representations for treatment effect estimation, IEEE Trans. Knowl. Data Eng. 35 (5) (2022) 4989–5001.
[22] S. Jiang, Q. Chen, Y. Xiang, Y. Pan, X. Wu, Y. Lin, Confounder balancing in adversarial domain adaptation for pre-trained large models fine-tuning, Neural Netw. (2024) 106173.
[23] J. Ma, M. Wan, L. Yang, J. Li, B. Hecht, J. Teevan, Learning causal effects on hypergraphs, in: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, 2022, pp. 1202–1212.
[24] P. Schwab, L. Linhardt, W. Karlen, Perfect match: A simple method for learning representations for counterfactual inference with neural networks, 2018, arXiv preprint arXiv:1810.00656.
[25] M. Cheng, X. Liao, Q. Liu, B. Ma, J. Xu, B. Zheng, Learning disentangled representations for counterfactual regression via mutual information minimization, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 1802–1806.
[26] Z. Ziyu, K. Kuang, F. Wu, Estimating treatment effect via differentiated confounder matching, in: Artificial Intelligence: First CAAI International Conference, CICAI 2021, Hangzhou, China, June 5–6, 2021, Proceedings, Part I 1, Springer, 2021, pp. 689–699.
[27] U. Shalit, F.D. Johansson, D. Sontag, Estimating individual treatment effect: generalization bounds and algorithms, in: International Conference on Machine Learning, PMLR, 2017, pp. 3076–3085.
[28] N. Hassanpour, R. Greiner, CounterFactual regression with importance sampling weights, in: IJCAI, 2019, pp. 5880–5887.
[29] L. Yao, S. Li, Y. Li, M. Huai, J. Gao, A. Zhang, Representation learning for treatment effect estimation from observational data, Adv. Neural Inf. Process. Syst. 31 (2018).
[30] S. Zeng, S. Assaad, C. Tao, S. Datta, L. Carin, F. Li, Double robust representation learning for counterfactual prediction, 2020, arXiv preprint arXiv:2010.07866.
[31] N. Hassanpour, R. Greiner, Learning disentangled representations for counterfactual regression, in: International Conference on Learning Representations, 2019.
[32] W. Zhang, L. Liu, J. Li, Treatment effect estimation with disentangled latent factors, in: Proceedings of the AAAI Conference on Artificial Intelligence, Vol. 35, 2021, pp. 10923–10930.
[33] J. Yuan, A. Wu, K. Kuang, B. Li, R. Wu, F. Wu, L. Lin, Auto iv: Counterfactual prediction via automatic instrumental variable decomposition, ACM Trans. Knowl. Discov. Data (TKDD) 16 (4) (2022) 1–20.
[34] L. Yao, Z. Chu, S. Li, Y. Li, J. Gao, A. Zhang, A survey on causal inference, ACM Trans. Knowl. Discov. Data (TKDD) 15 (5) (2021) 1–46.
[35] G. Maldonado, S. Greenland, Estimating causal effects, Int. J. Epidemiol. 31 (2) (2002) 422–429.
[36] Y. Bengio, A. Courville, P. Vincent, Representation learning: A review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[37] X. Wang, H. Chen, Z. Wu, W. Zhu, et al., Disentangled representation learning, IEEE Trans. Pattern Anal. Mach. Intell. (2024).
[38] M. Hernan, J. Robins, Causal Inference: What If, Chapman & Hill/CRC, Boca Raton, 2020.
[39] P. Cheng, W. Hao, S. Dai, J. Liu, Z. Gan, L. Carin, CLUB: A contrastive log-ratio upper bound of mutual information, 2020, arXiv preprint arXiv:2006.12013.
[40] G. Tesei, S. Giampanis, J. Shi, B. Norgeot, Learning end-to-end patient representations through self-supervised covariate balancing for causal treatment effect estimation, J. Biomed. Inform. 140 (2023) 104339.
[41] F. Johansson, U. Shalit, D. Sontag, Learning representations for counterfactual inference, in: International Conference on Machine Learning, PMLR, 2016, pp. 3020–3029.
[42] C. Shi, D. Blei, V. Veitch, Adapting neural networks for the estimation of treatment effects, Adv. Neural Inf. Process. Syst. 32 (2019).
[43] J.L. Hill, Bayesian nonparametric modeling for causal inference, J. Comput. Graph. Statist. 20 (1) (2011) 217–240.
[44] R.J. LaLonde, Evaluating the econometric evaluations of training programs with experimental data, Am. Econ. Rev. (1986) 604–620.

[45] J. Yoon, J. Jordon, M. Van Der Schaar, GANITE: Estimation of individualized treatment effects using generative adversarial nets, in: International Conference on Learning Representations, 2018.
[46] R. Guo, J. Li, H. Liu, Learning individual causal effects from networked observational data, in: Proceedings of the 13th International Conference on Web Search and Data Mining, 2020, pp. 232–240.

Dianlong You (Member, IEEE) received the Ph.D. degree in computer application technology from Yanshan University, China, in 2014. He is currently the professor and Ph.D. supervisor at the School of Information Science and Engineering, Yanshan University, China. From 2017 to 2018, he was a visiting scholar with the School of Computing and Informatics, University of Louisiana at Lafayette, Lafayette, Louisiana. From 2022 to 2023, he was a visiting scholar with the Data Science Institute, University of Technology Sydney, Sydney, NSW, Australia. His research interests include machine learning, streaming feature selection, and causal discovery. He has more than 20 publications including journals of IEEE Transactions on Neural Networks and Learning Systems, IEEE Transactions on Knowledge and Data Engineering, Information Fusion, Information Sciences, and Knowledge-Based Systems, etc.

Dongyan Wang currently is a Master Student in the School of Information Science and Engineering, Yanshan University, Qinhuangdao, HeBei. She current research interests include causal representation learning and treatment effect estimation.

Bingxin Liu is currently a Master Student in School of Information Science and Engineering, Yanshan University, Qinhuangdao, HeBei. She current research interests are focused on and causal inference.

Xiaoyi Ge is currently a Master Student in School of Information Science and Engineering, Yanshan University, Qinhuangdao, HeBei. She current research interests are focused on and causal inference.

Di Wu(Member, IEEE) received his Ph.D. degree from the Chongqing Institute of Green and Intelligent Technology (CIGIT), Chinese Academy of Sciences (CAS), China in 2019 and then joined CIGIT, CAS, China. He is currently a Professor of the College of Computer and Information Science, Southwest University, Chongqing, China. He has over 80 publications, including 21 IEEE TRANSACTIONS papers and several conference papers on AAAI, ICDM, WWW, IJCAI, etc. He is serving as an Associate Editor for the Neurocomputing and Frontiers in Neurorobotics. His research interests include machine learning and data mining. His homepage: https://wudi1986.github.io/Homepage/.

Xindong Wu received his Bachelor's and Master's degrees in Computer Science from the Hefei University of Technology, China, and his Ph.D. degree in Artificial Intelligence from the University of Edinburgh, Britain, in 1993. He currently is the Director and Professor of the Key Laboratory of Knowledge Engineering with Big Data (the Ministry of Education of China), Hefei University of Technology, China. He is a Foreign Member of the Russian Academy of Engineering and a Fellow of IEEE and the AAAS. He is the Steering Committee Chair of ICDM and the Editor-in-Chief of KAIS. His research interests include big data analytics, data mining, and knowledge engineering.